Privacy-Preserving Simulation of Facial Palsy Expressions Using Diffusion Models

Atsushi Tajima [1], Masataka Seo [2], and Yen-Wei Chen [1,*]

[1] Graduate School of Information Science and Engineering, Ritsumeikan University, Osaka, Japan
[2] Faculty of Robotics and Design, Osaka Institute of Technology, Osaka, Japan
Email: is0565xe@ed.ristumei.ac.jp (A.T.); masataka.seo@oit.ac.jp (M.S.); chen@is.ritsumei.ac.jp (Y.-W.C.)
*Corresponding author

***Abstract*—Facial nerve palsy results in impaired voluntary facial movements, and in Japan its severity is commonly assessed using the 40-point Yanagihara method. However, visual assessment is prone to inter-rater variability, and the use of actual patient images for educational purposes is restricted due to privacy concerns. To overcome these limitations, we propose a simulation framework that produces synthetic facial palsy images by leveraging a Stable Diffusion model with a partially fine-tuned ControlNet. Fine-tuning is confined to timesteps representing higher-level features, which allows the generation of pathological expressions while reducing the likelihood of reproducing patient-specific identity. Furthermore, a preprocessing stage modifies the size and placement of facial components and incorporates contour information from other faces. This design helps keep the fidelity of simulated expressions while maintaining strong privacy protection. The proposed method offers a practical resource for medical training and clinical research, where realistic and privacy-preserving facial data are required.**

***Keywords*—facial palsy, diffusion models, ControlNet, fine-tuning, privacy preservation**

I. Introduction

Facial nerve palsy is a condition that paralyzes the nerves controlling facial muscles, resulting in loss of voluntary movement in parts of the face. According to epidemiological data, the incidence is about 20–30 cases per 100,000 population [1]. Since the disorder may cause long-term impairment, the diagnosis requires selecting an appropriate treatment depending on the degree of paralysis. In Japan, the Yanagihara 40-point method is the most commonly used for this purpose [2]. In this assessment, the severity is scored based on 10 facial expressions, including resting asymmetry and nine voluntary movements. Fig. 1 presents the 10 expressions evaluated in the Yanagihara method. However, because this scoring is based on visual inspection, results may differ among physicians, particularly less experienced ones [3, 4]. To overcome this issue, educational tools that provide objective reference indicators are needed to unify evaluation standards. Nevertheless, due to privacy concerns, real patient facial images cannot be employed in academic meetings or medical training.

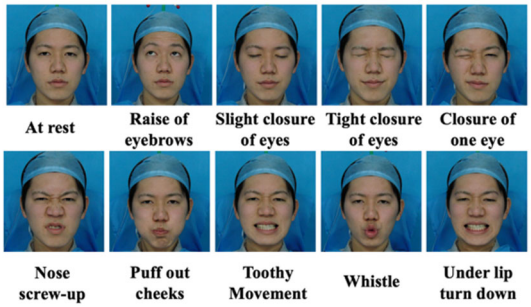

Fig. 1. The 10 facial expressions of the Yanagihara method.

To solve this privacy problem, our previous works employed Generative Adversarial Network (GAN)-based generative models to synthesize virtual facial images resembling the expressions of patients with facial palsy [5, 6]. However, due to the limited availability of patient data, the generated results were unsatisfactory, often inconsistent with the patient's identity and expressions, and accompanied by noise-like artifacts. To address these limitations, in this paper we propose a diffusion-based approach using Stable Diffusion with ControlNet. By introducing partial fine-tuning and expression-based conditioning, our method reduces artifacts and achieves more accurate clinical features while protecting patient privacy.

II. Literature Review

Previous studies on virtual facial image generation for facial palsy have mainly relied on GAN-based approaches, including our own prior works. These methods enabled the reproduction of pathological expressions but were limited by issues such as data scarcity, identity inconsistency, and the presence of artifacts. In parallel with these efforts, diffusion models have recently attracted considerable attention in the broader image generation field, achieving remarkable progress in fidelity and controllability. In the following, we first review GAN-based approaches for

Manuscript received October 17, 2025; revised October 28, 2025; accepted November 21, 2025; published April 28, 2026.

facial palsy simulation and then introduce diffusion-based models as a background for our proposed method.

A. GAN-Based Approaches for Facial Palsy Simulation

GAN-based approaches have been investigated for simulating facial palsy, and our previous works have contributed to this line of research. We have explored several GAN-based frameworks, including Multi-Condition Generative Adversarial Network (MC-GAN) [5, 6], pix2pix [7], and CycleGAN [8], with the aim of generating virtual facial images that preserve the identity of public faces while reproducing pathological expressions of patients. These studies represent initial attempts to utilize adversarial learning for clinical facial expression simulation, addressing the need for privacy-preserving yet clinically relevant data.

MC-GAN [9] is a conditional generative model that synthesizes images based on multiple input conditions. In our study, MC-GAN [5, 6] was adapted for facial palsy simulation by combining a public face image with a landmark-based expression image derived from patients, enabling the reproduction of characteristic pathological expressions. While this model demonstrated the feasibility of transferring expression features, it frequently produced unnatural appearances due to mismatches in facial shape between patients and public faces. Moreover, the scarcity of patient data further reduced the fidelity of generated expressions, limiting its applicability in practice.

Faceswap [10] is a face manipulation framework originally based on autoencoders, designed to swap facial identity while preserving attributes such as pose and expression. Although robust to differences in face shape, the model was limited in fidelity when training data were scarce. To improve this baseline, we incorporated adversarial learning into the Faceswap structure, introducing additional discriminators and loss functions, like the pix2pix framework [11]. This extension [8] enhanced identity preservation and image quality, but it still failed to accurately reproduce the symptomatic expressions of patients.

CycleGAN [12] further extended this approach by adapting the Faceswap structure with two generators, two discriminators, and a cycle consistency loss to enable unpaired image-to-image translation. This modification alleviated the dependence on paired training data and achieved higher fidelity and consistency compared with earlier GAN-based methods. Nevertheless, instability, artifacts, and insufficient reproduction of patient-specific expressions persisted.

Overall, these GAN-based approaches highlight both the potential and the limitations of adversarial learning in simulating facial palsy. Although they introduced strategies for privacy-preserving virtual face generation, challenges such as artifacts, unstable training dynamics, and incomplete reproduction of pathological expressions remain unsolved. These shortcomings motivate the exploration of alternative generative paradigms that can achieve higher stability, fidelity, and clinical reliability.

B. Denoising Diffusion Probabilistic Models (DDPM)

In recent years, diffusion models have achieved remarkable success in image generation, surpassing GANs and other generative frameworks in terms of fidelity, diversity, and training stability. These models learn to synthesize high-quality images by gradually reversing a noise perturbation process and have become the foundation for many state-of-the-art generative systems.

Denoising Diffusion Probabilistic Model (DDPM) [13] represents the most fundamental formulation of this framework. The model consists of two stages: a forward process and a reverse process. The forward process $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ gradually corrupts the input data $\boldsymbol{x}_0$ by adding Gaussian noise over a sequence of timesteps and is defined as a Markov chain. Here, ϵ denotes a noise sample drawn from a Gaussian distribution, and $\bar{\alpha}_t := \prod_{s=0}^{t} \alpha_s \,.\, \alpha_t = 1 - \beta_t$ is a differentiable function of timestep t. This formulation enables the input data distribution to be smoothly diffused into pure Gaussian noise as the number of steps increases. Intuitively, this process can be understood as progressively adding small amounts of noise to an image until all recognizable information is lost, similar to repeatedly blurring or degrading the image over time.

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon,\ \ \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \qquad (1)$$

The reverse process is the gradual removal of noise, in which a neural network is trained to reconstruct the original image by iteratively denoising the corrupted samples. At each timestep, the model learns to predict the noise that was added in the forward process, and by subtracting this predicted noise, the image distribution is progressively restored. In practice, this is implemented with a denoising U-Net operating directly in pixel space. In simple terms, the model learns to guess the amount of noise that was added at each step, so that it can later remove it in the reverse process. This learning is achieved by minimizing the difference between the true noise and the predicted one. The training objective is defined as the reconstruction loss between the predicted noise and the true noise, as follows.

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{x}_0,t,\epsilon\sim\mathcal{N}(0,1)}\left[\left\|\epsilon - \epsilon_\theta\left(\boldsymbol{x}_t, \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon\right)\right\|_2^2\right] \qquad (2)$$

During the inference, given a random noise $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, we can predict final denoised image $\boldsymbol{x}_0$ with the step-by-step reverse process. Where $\sigma_t = \frac{1-\bar{\alpha}_t}{\sqrt{1-\bar{\alpha}_t}}\beta_t$ is the variance of posterior Gaussian distribution $p_\theta(\boldsymbol{x}_0)$. Conceptually, this reverse process can be viewed as gradually revealing a clean image from random noise, similar to watching an image slowly emerge from static.

$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\boldsymbol{x}_t, t)\right) + \sigma_t \boldsymbol{z} \qquad (3)$$

C. Latent Diffusion Models (LDM)

Latent Diffusion Model (LDM) [14] is a type of diffusion model that combines a Denoising Diffusion

Probabilistic Model (DDPM) with an autoencoding framework. Instead of operating directly in the high-dimensional pixel space, the model performs the diffusion process in a compressed latent representation learned by a variational autoencoder, which is perceptually equivalent but computationally more efficient. This design drastically reduces memory and runtime requirements while maintaining visual quality.

In addition, LDM introduces cross-attention layers within the denoising U-Net between encoder and decoder blocks, enabling conditioning on various modalities such as text, segmentation maps, or edge images, which allows for flexible multi-modal image generation. The loss function is as follows. This setup allows the model to guide the denoising process according to the given condition c_t, ensuring that the generated result remains consistent with the provided guidance.

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{z}_0,t,c_t,\epsilon\sim\mathcal{N}(0,1)}[\|\epsilon - \epsilon_\theta(\boldsymbol{z}_t, t, c_t)\|_2^2] \quad (4)$$

Stable Diffusion [14] used in this study is a large-scale text-to-image diffusion system built on the LDM framework. It has been pretrained on massive image–text datasets, and its cross-attention mechanism aligns latent visual features with linguistic representations.

III. Materials and Methods

In this study, we propose a diffusion-based framework to generate virtual facial images that faithfully reproduce pathological expressions of facial palsy while preserving patient privacy. The overall framework of the proposed method is shown in Fig. 2. Our method is built upon Stable Diffusion with ControlNet, which serves as the backbone architecture for incorporating additional spatial conditions. To address the limitations of conventional fine-tuning, we restrict parameter updates to specific timesteps that correspond to high-level expression features, a strategy we refer to as partial fine-tuning. Moreover, to better reproduce clinical symptoms while avoiding resemblance to real patients, we design a conditioning mechanism that integrates contour guidance and organ-size deformation into the input representation.

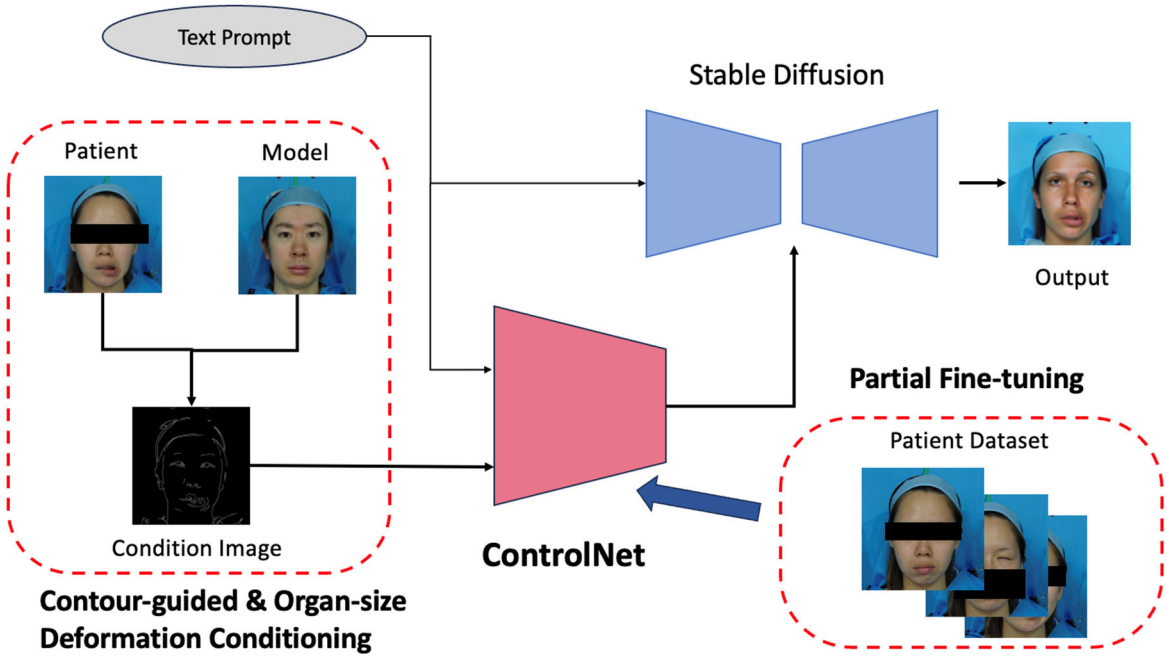

Fig. 2. Overall framework of the proposed method.

In the following subsections, we describe the details of each component: the backbone architecture (ControlNet), the fine-tuning strategy (partial fine-tuning), and the conditioning mechanism (contour-guided and organ-size deformation).

A. Backbone Architecture (ControlNet)

It is generally challenging to reproduce the subtle and complex facial expressions observed in facial paralysis when relying solely on text prompts as the conditioning input. To address this limitation, our study incorporates Stable Diffusion in combination with ControlNet [15]. ControlNet is a neural network framework specifically designed to extend large-scale pretrained text-to-image diffusion models with additional image-based conditioning signals, thereby enabling more precise guidance during generation.

As illustrated in Fig. 3, the architecture of ControlNet is constructed by duplicating the encoder weights of the base diffusion model and inserting zero-initialized convolutional layers around them. This design allows the base model to remain fixed during training while only the ControlNet parameters are updated, which significantly reduces computational cost. The zero-initialized layers play a stabilizing role at the start of training by suppressing spurious noise, ensuring that the model learns meaningful spatial correspondences rather than overfitting early.

The training process is guided by a noise prediction objective. The training loss is defined as follows, and it

ensures that ControlNet learns to utilize the conditional image input effectively during generation.

$$\mathcal{L} = \mathbb{E}_{z_0,t,c_t,c_f,\epsilon\sim\mathcal{N}(0,1)}[\|\epsilon - \epsilon_\theta(\boldsymbol{z}_t, t, c_t, c_f)\|_2^2] \quad (5)$$

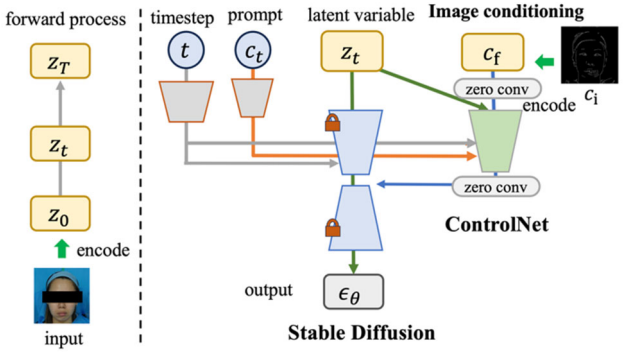

Fig. 3. Training of a ControlNet model.

As the conditioning input, we employ Canny edge maps. This choice is motivated by the fact that edge representations preserve the geometric structure of facial expressions while suppressing texture details associated with an individual's appearance.

In our experiments, we adopted a pretrained ControlNet model designed for Canny edge conditioning and further adapted it using our patient dataset. Although this fine-tuning improved the controllability of expression generation, the reproduced expressions were not perfectly faithful, and the resulting facial images often appeared slightly blurred.

B. *Fine-Tuning Strategy (Partial Fine-Tuning)*

In the reverse diffusion process, the generative model reconstructs an image in a coarse-to-fine manner, where global and large-scale structures are synthesized in the earlier timesteps, and progressively finer, more detailed features emerge in the later timesteps. Leveraging this property, we propose to restrict fine-tuning of the pretrained ControlNet to a limited range of timesteps that correspond to abstract and high-level representations. By focusing the training on this interval, the model is encouraged to generate facial images that accurately reproduce the disease expressions of patients while avoiding the reproduction of patient-specific appearance information.

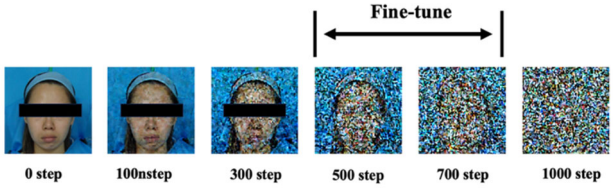

Fig. 4. The range of timesteps that we fine-tune the ControlNet.

Fig. 4 illustrates the selected range of timesteps used for partial fine-tuning. In our setting, the base diffusion model operates with the standard 1000-step schedule. Among these, we constrain fine-tuning to the interval $500 \leq t \leq 700$. Within this range, the generated noisy images primarily encode coarse expression-related structures, and the detailed identity-specific appearance features have not yet emerged. This makes the interval particularly suitable for learning expression-focused representations without overfitting to individual facial textures.

C. *Conditioning Mechanism (Contour-Guided & Organ-Size Deformation)*

To construct the conditioning images for our model, we designed a preprocessing pipeline, illustrated in Fig. 5. In the first step, facial landmarks are extracted from both the patient and another subject image. For landmark detection, we employ MediaPipe [16], an open-source framework developed by Google that is capable of detecting up to 468 facial landmarks in real time. These landmarks cover detailed regions such as the eyes, nose, and mouth, allowing us to precisely identify the areas relevant to expression changes. Based on these points, rectangular regions surrounding the eyes, nose, and mouth are cropped.

Next, the cropped regions are processed with an edge detector and resized to match the facial scale of the non-patient subject. In the final stage, the processed facial components are repositioned according to the edge contour of the non-patient subject's face.

Through this procedure, appearance-related cues are suppressed, while the geometric characteristics responsible for conveying pathological expressions are preserved. As a result, the constructed edge images differ substantially from the original patient's facial appearance, ensuring privacy protection, and at the same time provide ControlNet with effective structural conditions that enhance expression reproducibility during inference.

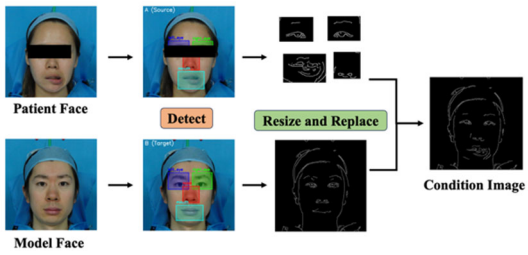

Fig. 5. Process of preprocessing a condition image.

IV. Result and Discussion

To verify the effectiveness of the proposed method, we conducted a series of experiments focusing on both implementation feasibility and performance evaluation. First, we describe the implementation details, including dataset configuration, preprocessing, and model settings. Next, we present the quantitative and qualitative results, where the proposed method is compared with baseline approaches in terms of expression fidelity, visual dissimilarity, and naturalness. Finally, we provide a discussion to interpret the outcomes, highlight the strengths of our approach, and examine remaining limitations and future directions.

A. *Implementation Details*

We used facial expression datasets provided by the Osaka Police Hospital for our experiments. The dataset

was provided through institutional collaboration in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (Ethics Committee) of Osaka Police Hospital (Reference Number: 1319). Personal identifiers were removed from all data, and the images were anonymized prior to use to ensure privacy protection. The dataset consists of facial expression images collected from three patients diagnosed with facial nerve palsy. For each patient, 10 distinct facial expressions were recorded, with 60 images per expression, resulting in a total of 600 images per patient and 1800 images overall. The dataset was carefully preprocessed to ensure consistency: face regions were cropped based on detected landmark points, and pixel values were normalized to a range of −1 to 1. All images were then resized to 512×512 pixels.

To generate conditioning inputs, we employed the corresponding Canny edge maps derived from the preprocessed images. In addition, for consistency during training, we assigned a fixed conditioning text prompt of *"ID photo of person wearing light-blue surgical cap and top, frontal headshot centered, even lighting, light-blue background, ultra-realistic, sharp focus"* to all images in the dataset. This design ensured that the generative model focused primarily on reproducing clinical features while suppressing patient-identifiable features. During the generation phase, the term "*person*" in the text prompt was replaced with different identities (e.g., "*man*," "*woman*," or specific appearance descriptors) so that the model could synthesize expressions on various non-patient faces while maintaining consistency in clothing and background.

The proposed model was fine-tuned on this dataset for two epochs with a batch size of 4. As initialization, we adopted the publicly available Stable Diffusion weights "*v*1-5-pruned-emaonly.ckpt" [17], pretrained on the large-scale LAION-2B-en dataset [18] consisting of approximately 2.3 billion image-text pairs. For the ControlNet component, we used the "control_sd15_canny.pth" weights [19], which are pretrained for edge-based conditioning. All experiments were implemented in Python using the PyTorch framework.

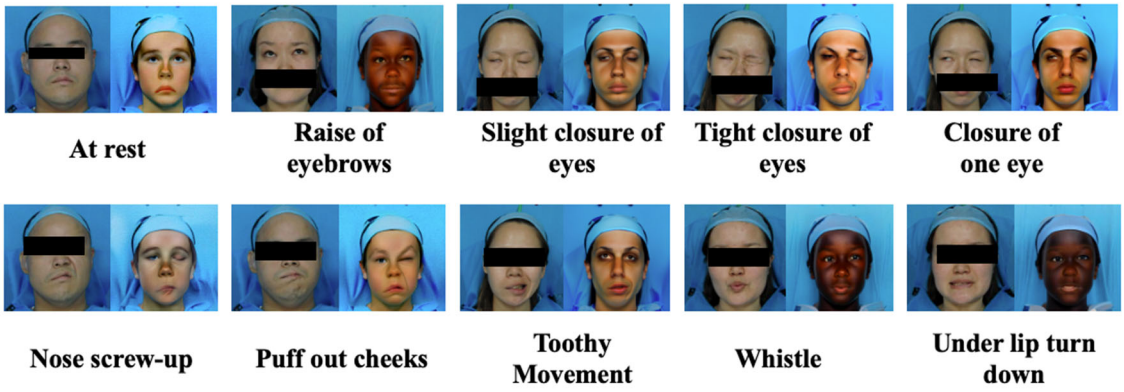

Fig. 6. Generated results for the ten facial expressions of the Yanagihara method. Left: patient images; Right: outputs from the proposed model.

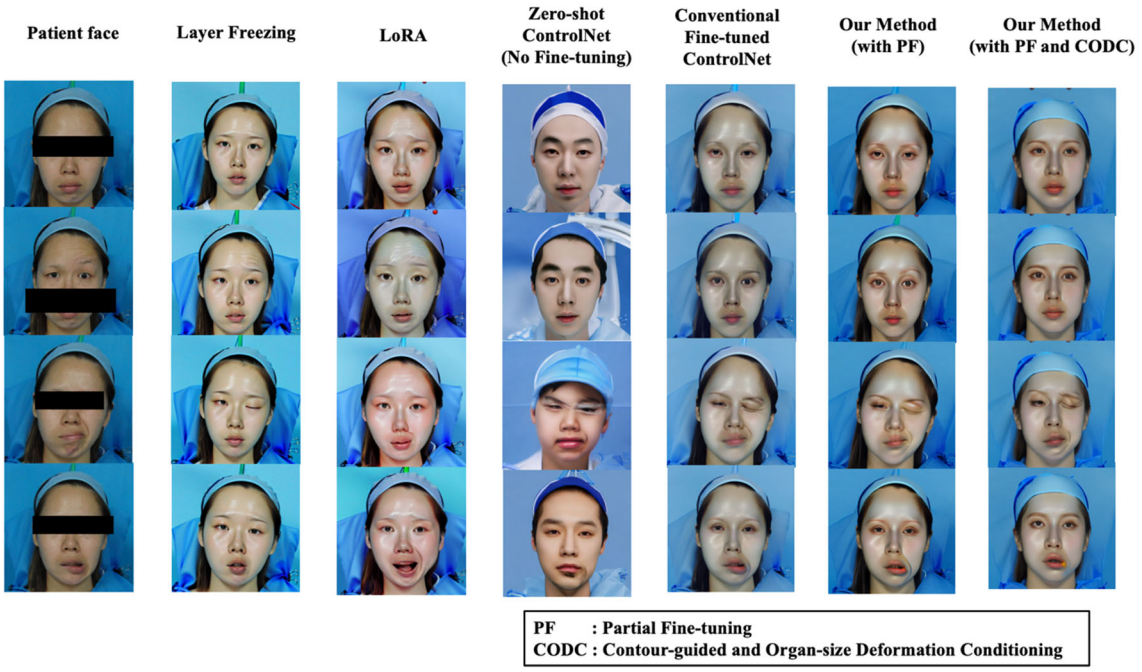

Fig. 7. Qualitative comparison with baseline models. From left to right: patient image, zero-shot model, fully fine-tuned model, Partial Fine-tuning (PF) (Ours), and Partial Fine-tuning (PF) with Contour-guided and Organ-size Deformation Conditioning (CODC) (Ours).

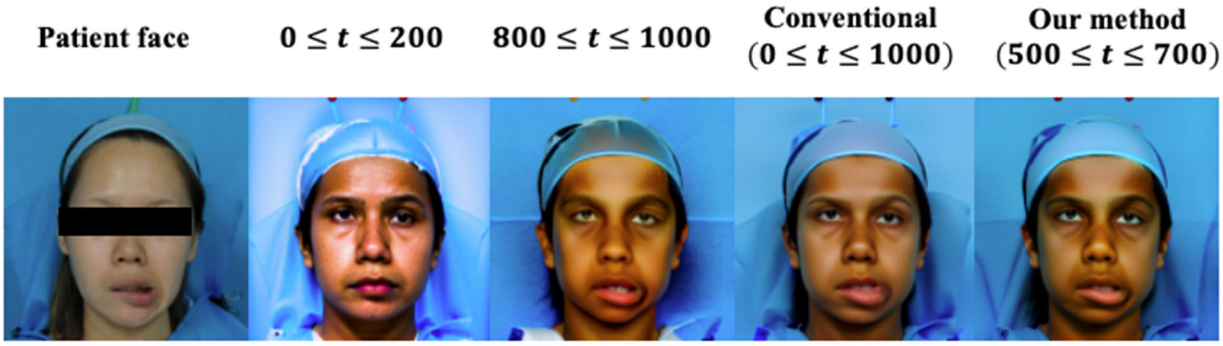

Fig. 8. Qualitative comparison under different timestep ranges. From left to right: patient image, 0–200 (late), 800–1000 (early), 0–1000 (full), and 500–700 (proposed).

B. *Quantitative and Qualitative Results*

To begin the qualitative evaluation, we first present the generated results for the ten facial expressions defined by the Yanagihara method. For each expression, the patient's image is shown on the left and the corresponding output generated by our proposed model is shown on the right in Fig. 6.

Following this, we compared our proposed approaches with four baseline models to assess expression reproducibility and privacy preservation: a model fine-tuned with Layer Freezing (all layers were frozen except the innermost blocks in the down, mid, and up paths, as well as the cross-attention layers) [20], a model fine-tuned using LoRA (Low-Rank Adaptation) [21], the zero-shot ControlNet model (pretrained ControlNet without fine-tuning), and a ControlNet model fine-tuned over the entire range of timesteps**.** Representative results are illustrated in Fig. 7. From left to right, the outputs correspond to the patient's face, the Layer Freezing model, the LoRA-based model, the zero-shot ControlNet, the fully fine-tuned ControlNet, Partial Fine-tuning only (Ours), and Partial Fine-tuning with Contour-guided and Organ-size Deformation Conditioning (Ours).

The images generated by the Layer Freezing and LoRA-based models exhibit structural instability, such as unnatural mouth openings and irregular eye closures, indicating insufficient preservation of local expression geometry. In addition, the zero-shot model, which has not undergone fine-tuning, generated nearly symmetrical faces and failed to capture pathological traits. Both of our proposed fine-tuned models reproduce asymmetrical expressions, but the model with partial fine-tuning more accurately represents fine details, such as the degree of elevation in the left eyebrow for the "Raise of Eyebrows" expression. Furthermore, the model with preprocessing produces facial images that differ in overall contour and in the relative balance of facial components such as the eyes, nose, and mouth, thereby enhancing visual dissimilarity from the original patient. However, we also observed that introducing the proposed conditioning slightly reduces expression reproducibility compared with partial fine-tuning alone.

In addition, we conducted an ablation study on the range of timesteps used for partial fine-tuning. Fig. 8 shows, from left to right, the patient's image, partial fine-tuning over 0–200, partial fine-tuning over 800–1000, the conventional fine-tuning across timesteps 0–1000, and the proposed method with 500–700. The results indicate that fine-tuning over 0–200 produces faces with nearly symmetrical expressions, showing little evidence of pathological traits. Fine-tuning over 800–1000 generates more asymmetric expressions but introduces visual artifacts, such as yellowing in the sclera (white regions of the eyes). The conventional full-range fine-tuning reproduces asymmetry but sometimes degrades local details, for example, generating distorted edges around the mouth. The proposed mid-range fine-tuning (500–700) achieves a better balance, reproducing expressions faithfully while avoiding the undesired artifacts observed in other ranges.

To further assess the performance of our method, we conducted a subjective evaluation through a questionnaire survey. Fourteen participants were asked to evaluate the generated facial expression images with respect to three criteria: expression reproducibility, visual dissimilarity from the patient's face, and naturalness as a facial image. Each criterion was rated on a five-point scale, where a higher score indicates better reproduction of expressions, greater dissimilarity from the original patient, and a more natural appearance. The average scores are summarized in Table I.

The results show that the model with partial fine-tuning alone achieved the highest scores in expression reproducibility and naturalness but was insufficient in terms of visual dissimilarity compared with other models. By incorporating the proposed conditioning mechanism, the visual dissimilarity was substantially improved, while the scores for expression reproducibility and naturalness decreased but still remained higher than those of the zero-shot model.

TABLE I. USER STUDY RESULTS (MEAN AND STANDARD DEVIATION)

Methods	PF	CODC	Expression Accuracy	Identity Diff	Naturalness
Zero-shot ControlNet			2.31/1.07	3.64/1.60	3.16/1.36
Conventional Fine-tuned ControlNet			3.79/0.96	3.01/1.17	3.95/0.83
Our Method 1	✓		**4.04/0.84**	3.01/1.29	**4.28/0.70**
Our Method 2	✓	✓	3.38/1.13	**3.80/1.32**	3.50/1.25

To objectively assess the quality and fidelity of generated facial images, we conducted a quantitative evaluation using three established metrics: Learned Perceptual Image Patch Similarity (LPIPS) [22], ArcFace similarity [23], and Face Image Quality Assessment via Stochastic Embedding Robustness (SER-FIQ) [24]. LPIPS evaluates perceptual quality by quantifying local structural and textural differences between each generated image and the corresponding patient's expression image, thereby assessing the fidelity of expression reproduction. ArcFace similarity, computed from the cosine distance between identity embeddings, measures the degree of identity dissimilarity between the generated image and the original patient, serving as an indicator of privacy preservation. SER-FIQ assesses the overall face image quality by estimating the robustness and consistency of facial embeddings under stochastic perturbations, reflecting the reliability of the generated facial representation.

Table II summarizes the quantitative results across all models, including both baseline and proposed approaches. Lower LPIPS and ArcFace similarity values indicate higher perceptual realism and stronger identity separation, while higher SER-FIQ values represent greater structural stability and overall image reliability.

TABLE II. QUANTITATIVE COMPARISON OF FACIAL EXPRESSION GENERATION MODELS

Methods	LPIPS ↓	ArcFace-based Facial Similarity ↓	SER-FIQ ↑
Layer Freezing	0.351	0.592	0.484
LoRA	0.353	0.501	0.351
Zero-shot ControlNet	0.456	**0.131**	0.490
Conventional Fine-tuned ControlNet	0.247	0.520	**0.596**
Our Method 1	**0.217**	0.417	0.560
Our Method 2	0.378	0.335	0.508

The Zero-shot ControlNet, which was not fine-tuned, achieved the highest LPIPS value (0.456), indicating that it could not sufficiently reproduce the target expressions. Although its ArcFace similarity was the lowest (0.131), the generated faces were often symmetrical or geometrically distorted, failing to represent the pathological asymmetry observed in the patient images. The Layer Freezing and LoRA-based models showed moderate LPIPS values (0.351 and 0.353, respectively) but relatively low SER-FIQ (0.484 and 0.351), indicating insufficient expression fidelity and unstable geometric consistency.

The conventional fine-tuned ControlNet, trained over all timesteps, exhibited improved perceptual and structural metrics (LPIPS = 0.247, SER-FIQ = 0.596). However, its ArcFace similarity remained high (0.520), suggesting that the model preserved patient-specific identity features.

In contrast, the proposed Partial Fine-tuning (Our Method 1) achieved the lowest LPIPS value (0.217), demonstrating the most accurate reproduction of subtle expression variations, and reduced ArcFace similarity to 0.417, which is lower than that of the fully fine-tuned model. This result indicates that Partial Fine-tuning enhanced expression reproducibility while suppressing identity leakage. The extended version, Partial Fine-tuning with Contour-guided and Organ-size Deformation Conditioning (Our Method 2), further reduced ArcFace similarity to 0.335, but LPIPS increased to 0.378. This result shows that while the conditioning improved privacy preservation, it also caused a rise in perceptual dissimilarity. A moderate SER-FIQ score (0.508) confirmed that the generated faces maintained structural consistency despite this trade-off.

These results indicate that the proposed Partial Fine-tuning method improved expression reproducibility and privacy preservation compared with the baseline models, and that incorporating Contour-guided and Organ-size Deformation Conditioning enhanced privacy preservation.

C. Discussion

Our investigation demonstrates that the proposed method is effective in simulating pathological facial expressions while preserving patient privacy. In particular, the ablation study on timestep ranges revealed that partial fine-tuning within the mid-range (500–700) achieves the best trade-off between expression fidelity and identity suppression. Fine-tuning in the early range (800–1000) often produced unwanted artifacts such as discoloration in the sclera, while fine-tuning in the late range (0–200) failed to reproduce pathological traits, yielding nearly symmetrical outputs. These findings highlight the importance of selecting an appropriate timestep interval for expression-focused training.

The proposed contour-guided and organ-size deformation conditioning alters the structural prior of the generated face by incorporating non-patient contour information. This effectively suppresses the direct perceptual association with the patient's identity. However, replacing the original boundary information with an external prior also attenuates the constraints that preserve subtle pathological deformations, leading to a moderate reduction in expression reproducibility compared with partial fine-tuning alone.

The subjective evaluation further supports these findings. Fourteen participants assessed the generated results in terms of expression reproducibility, visual dissimilarity, and naturalness. The proposed method achieved higher scores than the zero-shot model in all categories, particularly in terms of dissimilarity from the patient. However, the scores for expression reproducibility and naturalness were lower than those of the partial fine-tuning only model. These results suggest that the proposed conditioning is effective for anonymization while maintaining sufficient fidelity and realism for educational applications.

Recent studies have also applied diffusion-based models to medical image generation, particularly in radiology and dermatology, demonstrating the potential of these models for synthesizing realistic and privacy-preserving medical data. Compared with these works, our approach targets facial palsy, a domain where facial structure and expression must be preserved simultaneously. This introduces unique challenges distinct from organ- or texture-level synthesis in prior medical domains. Our

results highlight that diffusion-based generation can also be effectively adapted to pathological facial expressions, expanding the applicability of such models beyond conventional medical imaging modalities.

Future work will focus on refining training strategies and conditioning mechanisms to further enhance visual dissimilarity from the patient while preserving pathological expressions. In addition, we plan to expand the dataset by including a larger and more diverse set of patients to improve the generalizability and robustness of the proposed model. Beyond still images, we also plan to extend the framework to generate pseudo-facial palsy expression videos, enabling the simulation of temporal dynamics that are valuable for clinical education and training.

V. Conclusion

In this study, we proposed a diffusion-based framework for simulating facial palsy expressions that balances expression fidelity and privacy preservation. By partially fine-tuning ControlNet within a limited range of timesteps, the model effectively reproduced pathological asymmetry. Furthermore, the introduction of contour-guided and organ-size deformation conditioning suppressed identity cues and enhanced visual dissimilarity from the patient, but reduced expression reproducibility compared with partial fine-tuning alone. Subjective evaluations confirmed that the proposed method outperformed the zero-shot baseline in all aspects, demonstrating its potential as an anonymized alternative for clinical education and training.

Overall, our findings highlight the suitability of diffusion-based approaches for medical image simulation, providing a promising direction for generating realistic yet privacy-preserving representations of pathological facial expressions. Future work will refine training strategies and conditioning mechanisms to further improve the balance between fidelity and dissimilarity, and extend the framework toward generating pseudo-facial palsy expression videos that capture temporal dynamics for broader clinical applicability.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Atsushi Tajima proposed the novel AI method, designed the framework, implemented the experiments, analyzed the data, and drafted the manuscript. Masataka Seo provided guidance on the research design, supervised the implementation and data analysis, and contributed to revising the manuscript. Yen-Wei Chen supervised the overall project, secured funding, and provided critical feedback on the manuscript. All authors discussed the results and approved the final version of the manuscript.

Acknowledgment

We are deeply grateful to Dr. Naoki Matsushiro of Osaka Police Hospital for his support during the initial stage of this study and for preparing the facial image database of facial nerve palsy patients and healthy individuals that was utilized in this work.

References

[1] K. K. Adour *et al.*, "The true nature of Bell's palsy: Analysis of 1000 consecutive patients," *Laryngoscope*, vol. 88, pp. 787–801, 1978.

[2] N. Yanagihara *et al.*, "A study on the criteria for determining the degree of facial neuropathy," *Journal of the Japanese Society of Otolaryngology*, vol. 80, no. 8, pp. 799–805, 1997.

[3] N. Matsushiro, "Differences in the evaluation of facial palsy (Yanagihara grading system) among 52 ENT doctors at Osaka University," *Facial Nerve Research*, vol. 29, pp. 60–62, 2010.

[4] N. Matsushiro, "Differences in the evaluation of facial palsy (Yanagihara Grading System) among 9 facial palsy specialists and 47 general ENT doctors in Japan: A 9 University collaborative investigation," *Facial Nerve Research*, vol. 29, pp. 63–65, 2010.

[5] S. Yaotome, M. Seo, N. Matsushiro, and Y.-W. Chen, "Simulation of facial palsy using conditional generative adversarial networks," in *Proc. 2020 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, 2010, pp. 579–582.

[6] T. Sakai, M. Seo, N. Matsushiro, and Y.-W. Chen, "Simulation of facial palsy using conditional generative adversarial networks and face shape normalization," in *Proc. 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, 2021, pp. 793–797.

[7] A. Tajima, M. Seo, N. Matsushiro, and Y.-W. Chen, "Development of a facial palsy expression simulator using faceswap with skip connections and patch discriminator," in *Proc. Joint Convention Record of the Kansai Chapters of Electrical Societies*, 2023.

[8] T. Sakai, M. Seo, N. Matsushiro, and Y.-W. Chen, "Simulation of facial palsy using an improved cycle GAN and face restoration network," in *Proc. 2023 IEEE Int. Conf. Consumer Electronics (ICCE),* 2023, pp. 1–4.

[9] H. Park, Y. Yoo, and N. Kwak, "MC-GAN: Multi-conditional generative adversarial network for image synthesis," in *Proc. British Machine Vision Conference (BMVC)*, 2018.

[10] Faceswap Developers. (2017). Faceswap: Open Source deepfake software [Online]. Available: https://faceswap.dev

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976.

[12] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 2223–2232.

[13] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 6840–6851, 2020.

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10684–10695.

[15] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2023, pp. 3836–3847.

[16] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "MediaPipe: A framework for building perception Pipelines," in *Proc. 28th ACM International Conference on Multimedia (ACM MM'20)*, 2020, pp. 835–844.

[17] Stable diffusion v1-5 model card. *Hugging Face*. [Online]. Available: https://huggingface.co/runwayml/stable-diffusion-v1-5

[18] C. Schuhmann *et al.*, "LAION-5B: An open large-scale dataset for training next generation image-text models," arXiv Preprint, arXiv:2210.08402, 2022.

[19] The pretrained weights of ControlNet. *Hugging Face*. [Online]. Available: https://huggingface.co/lllyasviel/ControlNet/tree/main/models

[20] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. 28th Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Montreal, QC, Canada, 2014, pp. 3320–3328.

[21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. 10th Int. Conf. Learn. Represent. (ICLR)*, Virtual Event, 2022.

[22] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 586–595.

[23] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit (CVPR)*, Long Beach, CA, USA, 2019, pp. 4690–4699.

[24] P. Terhörst, M. Huber, C. Herrmann, N. Damer, F. Kirchbuchner, and A. Kuijper, "SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit (CVPR)*, Seattle, WA, USA, 2020, pp. 5651–5660.