DiffusionCleft: Facial Anomaly Synthesis Guided by Text

Karen Rosero
kroseroj@andrew.cmu.edu
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA

Lucas M. Harrison
lucas.harrison@utsouthwestern.edu
Department of Plastic Surgery
UT Southwestern Medical Center
Dallas, Texas, USA

Alex A. Kane
alex.kane@utsouthwestern.edu
Department of Plastic Surgery
UT Southwestern Medical Center
Dallas, Texas, USA

Rami R. Hallac
rami.hallac@childrens.com
Dept. of Plastic Surgery, Analytical Imaging and Modeling Center
UT Southwestern Medical Center, Children's Health Dallas
Dallas, Texas, USA

Carlos Busso
busso@cmu.edu
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA

Abstract

Text-to-image diffusion models demonstrate strong capabilities in generating photorealistic content across diverse domains. However, they remain limited in synthesizing clinically relevant facial anomalies, such as cleft lip, due to the lack of domain-specific representations and adaptation strategies. In this work, we introduce a method for domain-specialized image generation by adapting a publicly available multimodal diffusion model to synthesize prompt-based, realistic facial images of both pre-operative and post-operative cleft lip conditioned on a small set of real images. We compute quantitative metrics to evaluate the realism, identity safety, and diversity of the generated images, including *face identity recognition* (FIR), *Fréchet inception distance* (FID), and *learned perceptual image patch similarity* (LPIPS). In addition, two medical experts independently rated a subset of the generated samples for anatomical plausibility and visual realism. Results show that the adapted model avoids identity leakage, outperforms previous GAN-based approaches in distributional similarity, and achieves average human ratings of 4.85 for realism and 4.81 for anatomical plausibility on a 5-point Likert scale. Beyond qualitative generation, we demonstrate the clinical utility of the generated images by training a lip anomaly detection model on synthetic samples, achieving an accuracy of 79% on real clinical data. These findings establish a new paradigm for adapting generative models toward generating diverse, clinically meaningful imagery with high fidelity and domain specificity.

CCS Concepts

• **Applied computing** → **Health care information systems**; **Imaging**.

ICMI '25, Canberra, ACT, Australia

ACM ISBN 979-8-4007-1499-3/2025/10
https://doi.org/10.1145/3716553.3750794

Keywords

Medical Imaging; Text-to-Image Generation; Diffusion Models; Cleft Lip; AI Methods for Healthcare; Privacy Preservation Strategy.

ACM Reference Format:
Karen Rosero, Lucas M. Harrison, Alex A. Kane, Rami R. Hallac, and Carlos Busso. 2025. DiffusionCleft: Facial Anomaly Synthesis Guided by Text. In *Proceedings of the 27th International Conference on Multimodal Interaction (ICMI '25), October 13–17, 2025, Canberra, ACT, Australia.* ACM, Canberra, AUS, 10 pages. https://doi.org/10.1145/3716553.3750794

1 Introduction

Cleft lip is a congenital craniofacial condition that occurs when the upper lip fails to fuse during fetal development. It is one of the most common birth defects worldwide, affecting 1 in 600 to 800 live births [23, 34, 38]. The condition can range in severity, from a small notch in the lip to a complete separation extending into the nose [35]. While cleft lip can occur in isolation, it is often associated with cleft palate and may lead to functional challenges such as difficulties with feeding, speaking, and hearing [4]. Surgical repair is typically performed within the first year of life and plays a critical role in restoring both function and facial appearance [9]. However, even after surgical intervention, subtle anatomical differences may persist, which can vary depending on the technique used, timing of the intervention, and individual healing responses [8].

Despite the importance of visual documentation for diagnosis, treatment planning, and outcome evaluation in cleft lip cases, collecting facial images of affected individuals presents several challenges. Medical facial datasets are inherently sensitive due to the identifiable nature of facial features and the associated stigma around visible differences [39]. As a result, institutions often face strict ethical and legal constraints regarding image sharing and publication, especially since the subjects are often minors. Patient consent, data anonymization, and privacy protection protocols further limit the availability and utility of such datasets for large-scale research and machine learning applications [7]. These constraints create a bottleneck for developing and benchmarking automated facial analysis systems in clinical contexts. Some restrictions also apply to new tools for image generation. Despite the recent advances in prompt-based image generation that offer photorealistic

outputs, the companies developing these tools enforce strong safety constraints[1]. For instance, prompts that include medical terms – such as cleft lip – are currently blocked due to content policy restrictions. While these safeguards are essential, they limit the use of general-purpose generative models for clinical research, especially in areas where real-world data is scarce.

Deep learning models have shown great promise in medical image analysis, including facial landmark detection [25, 30], diagnosis [36] and surgical outcome evaluation [31, 32]. However, these models rely on large, diverse, and well-labeled datasets to generalize effectively. In the context of cleft lip, the limited availability of curated facial image datasets restricts the performance and robustness of such models. This data scarcity is particularly acute for specific subgroups, such as patients with cleft lip, leading to potential biases and reduced clinical relevance. As a result, there is a growing need for synthetic data generation techniques that can augment existing datasets while preserving anatomical realism and variability, without compromising patient privacy.

The contribution of this work is a diffusion-based model that enables controlled generation of facial images depicting *pre-operatory* (pre-op) and *post-operatory* (post-op) cleft lip conditions using text-based prompts. We also refer to these conditions as *unrepaired* and *repaired* cleft lip, respectively. We build upon a state-of-the-art text-to-image model and adapt it to this medical task by fine-tuning on a small dataset of real facial images of individuals with cleft lip. Each training image was manually annotated with descriptive prompts that captured both the surgical status and the cleft subtype. By aligning text descriptions with visual features during fine-tuning, our approach enables prompt-based generation of clinically meaningful face images that reflect diverse anatomical structures.

Our experimental results demonstrate that the fine-tuned model does not memorize or reproduce identities from the training set, as confirmed by low *face identity recognition* (FIR) scores. The generated images achieve a favorable *Fréchet Inception Distance* (FID), indicating that they lie within the distribution of real cleft lip images. To assess diversity, we compute the *learned perceptual image patch similarity* (LPIPS) between image pairs, confirming high anatomical variability across generations. Additionally, two medical practitioners with expertise in cleft lip evaluated a subset of images for both realism and *anatomical plausibility* (AP), providing further validation of the clinical relevance of the generated samples. The synthesized images demonstrated practical utility by being used to train a lip anomaly detection system, which achieved 79% accuracy on real images—surpassing the 66% accuracy reported in comparable work [31]. Our method offers a new pathway for generating high-quality, privacy-preserving synthetic data to support training and evaluation in facial analysis tasks specific to cleft lip. The strategy is general and can be applied to other medical tasks.

2 Related Work

There has been growing interest on facial analysis of individuals with cleft lip, particularly through image inpainting approaches to synthesize post-operative images from pre-operative inputs [3, 12, 31]. However, these models rely on generative backbones pre-trained exclusively on control individuals included in these facial datasets, limiting their ability to synthesize anatomical irregularities characteristic of cleft-lip.In contrast, our work focuses on generating entirely new facial identities that depict either pre-op cleft lip morphology or clinically realistic post-op features, such as nasal and upper lip asymmetry, and philtrum scarring, rather than reconstructing or inpainting over existing subjects.

Aligned to our goal, GestaltGAN [19] generated photorealistic portraits of individuals with rare genetic disorders using a modified StyleGAN3-R architecture [15]. However, cleft lip was not included among the conditions studied. CleftGAN [11] also relied on StyleGAN3-R model for generating faces with cleft lip conditions. The model was initialized with weights from high-resolution images of faces from control individuals. Then, it was fine-tuned on a dataset of 514 cleft-affected frontal photographs. The resulting model achieved an FID of 21.03 on 2,000 generated samples. While the generated faces captured features relevant to cleft morphology, the images often exhibited background artifacts and blurring, which reduced their overall photorealism.

Current uni-modal diffusion-based approaches in the clinical domain have predominantly targeted specific imaging modalities, including MRI and CT scans [17], radiographs [18], dermoscopies for melanoma detection [1], and facial images of individuals with facial paralysis [10]. However, these methods have not leveraged the potential of cross-modal interactions between text and images. In this paper, we leverage the state-of-the-art capabilities of a pre-trained diffusion model for text-guided generation of facial images, with a specific focus on pre- and post-operative cleft lip cases.

3 Methodology

This section describes the proposed approach for adapting a pre-trained text-to-image diffusion model to generate facial images depicting cleft lip conditions. We outline the architecture of the base model, the fine-tuning procedure, and the preprocessing strategy used to prepare latent representations for training.

3.1 Base Model

Our approach builds on the FLUX.1 [schnell] model, developed by Black Forest Labs. This open-source text-to-image generation model integrates a hybrid architecture combining diffusion processes with transformer-based components [20]. It maintains separate parameter sets for language and image modalities, allowing effective integration of textual prompts into the image synthesis process. A distinguishing feature of FLUX.1 compared to earlier diffusion-based models, such as Stable Diffusion [26, 27], is its use of *flow matching* [22]. Rather than relying on traditional iterative denoising, flow matching learns a continuous trajectory from noise to image space, inspired by the framework of *continuous normalizing flows* [22]. This formulation allows for faster inference by reducing the number of denoising steps, without reducing realism.

Text conditioning is handled by a dual-encoder system comprising *contrastive language-image pre-training* (CLIP) [28] and T5 [29] models. The CLIP encoder captures visual-semantic relationships to align prompts with image features, while T5 is used to process more linguistically complex prompts. The image generation backbone consists of a UNet-like structure augmented with cross-modal attention layers incorporating features from both encoders. FLUX.1

[1] https://openai.com/index/introducing-4o-image-generation/

comprises approximately 12 billion parameters. Although the specific training data has not been disclosed, these models are typically trained on large-scale image-text pairs. Figure 1 shows images generated by different models before adaptation using the prompt *A boy's face with a cleft lip*. The pretrained model does not naturally capture the anatomical features associated with cleft lip, indicating the need for domain-specific adaptation. In some cases (Fig. 1d), it adds irrelevant artifacts to the face.

3.2 Fine-Tuning for Cleft Lip Image Generation

We adapt FLUX.1 [schnell] through parameter-efficient fine-tuning using *low-rank adaptation* (LoRA) [14]. Trainable low-rank projection layers are inserted into the UNet, while the original text encoders are kept frozen. This decision assumes that cleft-related terminology is already represented in the pretrained language embeddings and that the primary adaptation required is in the visual domain and its cross-modality representations.

All training images are grouped into resolution buckets based on aspect ratio and resized to one of four fixed resolutions (256, 512, 768, or 1024 pixels). This bucketing strategy introduces variability while preserving anatomical proportions. Each image is passed through a pretrained *variational autoencoder* (VAE), also included in the base model, to obtain a compressed latent representation. These latents are precomputed and cached to disk, reducing GPU memory usage during training and eliminating redundant computation.

We add a unique trigger word to the beginning of every training prompt to ensure that the model learns to generate cleft lip images only when intended. This word is expected to be an uncommon combination of characters not likely to appear in regular prompts used by the base model, serving as a special signal that activates the fine-tuned behavior. Using a trigger word ensures that the model focuses on learning cleft-specific features without affecting its ability to generate general images. During inference, the cleft lip characteristics will only appear in generated images if the trigger word is included in the prompt.

The model operates through rich multimodal interaction between text and image representations that align prompt semantics and facial anatomy in the latent space. Regular sampling is performed during training to qualitatively monitor prompt adherence and anatomical plausibility in the generated images. Upon completion of training, only the LoRA weights are saved, resulting in a lightweight adapter file that can be merged with or applied to the base model at inference time.

3.3 Prompt Design and Inference Strategy

To accommodate the anatomical and morphological differences between cleft lip presentations before and after surgical intervention, we train two separate LoRA-adapted diffusion models for pre- and post-op cleft lip. This decision allows each model to specialize in the visual and structural patterns specific to its respective class, ensuring accurate synthesis of class-relevant features. Pre-operative cleft lip is typically characterized by a visible tissue discontinuity in the upper lip, which may extend to adjacent regions such as the gingiva, hard or soft palate, or nasal cavity. These clefts can vary in severity and presentation, appearing as unilateral or bilateral conditions. In contrast, post-operative cleft lip reflects the outcomes of surgical repair. It may exhibit characteristics such as nasal asymmetry, an irregular or raised cupid's bow, lip asymmetry, or visible scarring in the upper lip or philtrum. These postoperative features depend on both the cleft severity and the surgical outcome.

Table 1: Prompt components used for image generation of pre- and post-op cleft lip.

	Pre-Op Cleft Lip	Post-Op Cleft Lip
Trigger	`CLP`	`CLPrep`
Age/gender	girl, boy, baby, kid	girl, boy, baby, kid, man, woman
Skin tone	light, pale, medium, fair, dark, olive	
Hair	curly, straight, short, wavy, long, blonde, brown, black, red	
Eye color	brown, green, blue, hazel	
Cleft features	unilateral or bilateral	subtle scar(s), slight lip/nose asymmetry
Status	unrepaired cleft lip	after cleft lip repair

Each model is queried through inference using prompts that contain a task-specific trigger word, which is an uncommon, non-semantic sequence. We use `CLP` for pre-operative and `CLPrep` for post-operative cleft lip synthesis. This strategy ensures that cleft-specific features are generated only when intended, while preserving the model's generalization. Prompts are constructed using structured combinations of attributes detailed in Table 1, including subject age group and gender, skin tone, hair characteristics, eye color, cleft type, and cleft status. Every prompt begins with the designated trigger word followed by a phrase describing the subject's full face and incorporates at least one cleft-relevant anatomical descriptor. An example of a prompt is:

[CLP] A girl's whole face with an unrepaired unilateral cleft lip. Light skin. Brown eyes.

Additionally, all prompts include the phrase *"Only one face, no text, no watermarks, no masks or occlusions"* at the end to minimize the generation of artifacts. We use this controlled prompting strategy to generate 2,000 images per class (pre- and post-op cleft lip). Prompts are sampled to encourage diversity while maintaining consistency across clinical and demographic attributes. The resulting image sets serve as the basis for the quantitative and qualitative evaluations presented in the subsequent sections.

4 Experimental Setup

This section describes the data, the training configuration used to fine-tune the FLUX.1 [schnell] model, and the metrics used to evaluate the image generation depicting pre- and post-op cleft lip.

4.1 Datasets

We collected 301 facial images of individuals affected by cleft lip. This dataset includes 100 images depicting pre-operative cleft lip and 201 images of patients after undergoing surgical repair. Images were sourced from publicly available online resources, including websites of cleft-specialized surgeons, hospitals, and non-profit

organizations that do not prohibit their use. Images include frontal poses in controlled and natural illumination. No training data samples are shown in this manuscript to preserve patient privacy.

For the training prompts, we avoided detailed identity-specific attributes (e.g., skin tone, eye color, or background) to reduce overfitting and support generalization. Instead, the captions focused on clinically relevant features, including age group (baby, child, adult), gender, cleft type (unilateral or bilateral), nasal asymmetry, repair status, and the presence of scarring in the orofacial region.

4.2 Model Hyperparameters

The finetuning process was initialized with the publicly available checkpoint on Hugging Face[2]. We implement LoRA adapters set to a rank of 16 and a scaling factor of 32 into the UNet backbone responsible for image synthesis, while the text encoder was kept frozen. Each facial image was paired with a corresponding caption containing the trigger word, `CLP` for pre-operative and `CLPrep` for post-operative cleft lip, to explicitly guide the model to generate features specific to the target domain. All images were preprocessed and resized into buckets ranging from 256 to 1024 pixels based on aspect ratio, and cached as latent representations.

The model was trained using a learning rate of 1×10^{-4} with the AdamW8bit optimizer. A batch size of 1 was used with no gradient accumulation. Sampling occurred at regular intervals using a resolution of 768×768, 4 denoising steps, and the maximum text guidance scale of 1. Prompt variations followed the description in Section 3.3.

We experimented finetuning on two hardware configurations: (1) an NVIDIA RTX 4090 GPU with 24GB of VRAM, where CPU offloading was required to manage memory usage, and (2) a Tesla A100 GPU with 40GB of VRAM, where offloading was unnecessary. The training batch size was fixed to 1, meaning each model iteration processes one image-caption pair. The total number of training iterations was set to $10N$, where N corresponds to the number of images in the dataset, ensuring that each sample was seen multiple times during training.

4.3 Metrics

We aim to generate diverse images with precise cleft-left anomalies. We also want our generated images to be different from the samples in the database. We employ a combination of quantitative and perceptual metrics used in generative modeling to evaluate the quality, realism, and diversity of the generated facial images

4.3.1 Evaluation of Face Re-Identification To verify that the model does not memorize or replicate identities from the training set, we calculate *face identity recognition* (FIR) based on embedding similarity. This evaluation is especially important in clinical applications, where privacy and ethical considerations prohibit identity leakage. We use the InsightFace framework[3], specifically the pre-trained `w600k_r50` model based on a ResNet-50 backbone trained on the WebFace600K dataset [6]. This model produces 512-dimensional face embeddings that capture identity-specific features. Cosine similarity is computed between the embeddings of each generated image and all real training images from the corresponding cleft category (pre- or post-operative). The analysis is performed on a pool of 2,000 generated samples per class. Since a similarity value closer to zero indicates a low likelihood of re-identification, we have used a threshold of 0.6 to classify a pair as having the same identity. This approach has been widely adopted in generative face modeling studies to assess identity leakage. It ensures that generated samples represent novel identities rather than reproductions of training examples [2, 5, 21].

4.3.2 Distributional Similarity Evaluation We compute the *Fréchet Inception Distance* (FID) [13] to assess the overall realism and distributional similarity between real and generated images. FID quantifies the distance between two multivariate Gaussians fitted to the feature embeddings extracted from a deep classification network – in our case, the Inception v3 model pre-trained on ImageNet-1k. One distribution is computed over real (training) images and the other over generated samples. The FID metric captures both mean and covariance differences between the two distributions. Lower FID values indicate greater alignment between real and synthetic data, with values closer to zero suggesting higher visual fidelity and realism. In this study, we compute FID separately for pre-operative and post-operative cleft lip images, using the PyTorch implementation[4] that aligns with the original TensorFlow version[5].

4.3.3 Diversity Evaluation We compute the *Learned Perceptual Image Patch Similarity* (LPIPS) [40] to measure intra-set perceptual diversity and ensure the model avoids mode collapse. LPIPS measures perceptual distances between image pairs by comparing deep features extracted from a pre-trained neural network. Unlike pixel-wise metrics, LPIPS is sensitive to structural and textural differences, making it more aligned with human visual perception. We compute LPIPS scores between 2,000 randomly selected pairs of generated images within each cleft category to quantitatively assess the level of structural variability present among the synthesized samples. This analysis is conducted using the official LPIPS implementation[6]. LPIPS scores closer to 1 indicate greater perceptual dissimilarity which may reflect excessive randomness, whereas scores closer to 0 may signal repetitive outputs. In the context of this work, LPIPS values closer to 0.5 are considered more favorable, as they indicate a balanced level of diversity. LPIPS complements FID by evaluating the model's capacity to generate diverse facial features within each class.

4.3.4 Human Evaluation of Realism and Anatomical Plausibility Quantitative metrics alone cannot capture clinical relevance or anatomical plausibility. We conducted a perceptual study with two medical practitioners experienced in cleft lip diagnosis and treatment to complement objective evaluations. A total of 50 generated images were independently rated by both experts using a 5-point Likert scale. Each image was scored on two criteria: overall visual realism, defined as how convincing the image is, and anatomical plausibility, defined as the feasibility of the generated facial anatomy from a medical perspective. A rating of 5 represents the highest level of realism or plausibility, while 1 indicates low quality or unlikely anatomical configuration. This human evaluation provides

[2] https://huggingface.co/black-forest-labs/FLUX.1-schnell
[3] https://github.com/deepinsight/insightface
[4] https://github.com/mseitzer/pytorch-fid/tree/master
[5] https://github.com/bioinf-jku/TTUR
[6] https://github.com/richzhang/PerceptualSimilarity

(a) SDXL model

(b) Phoenix by Leonardo.Ai

(c) FLUX 1.1 Pro model

(d) FLUX.1 [schnell] model

Figure 1: Images generated by different diffusion-based models for the prompt: *"A boy's face with cleft lip"* before model adaptation. These state-of-the-art models cannot generate cleft lip images

an additional layer of validation to assess whether the generated images could plausibly represent real-world cleft lip cases.

5 Experiments and Results

This section analyzes the results obtained for the fine-tuned models and interprets the quantitative and qualitative metrics considered for evaluation.

5.1 Inference on Zero-Shot Models

We began by investigating the zero-shot capabilities of several state-of-the-art text-to-image generation models in synthesizing facial images depicting a cleft lip without any domain-specific fine-tuning. For this purpose, we employed a straightforward prompt: *"A boy's face with a cleft lip"*. Figure 1 presents the output of diffusion-based models currently considered cutting-edge image generators.

Figure 1a was generated using SDXL, a publicly available latent diffusion model designed to produce photorealistic images with strong alignment to text prompts [27]. Figure 1b shows the output from the Phoenix model by Leonardo.Ai, a commercial diffusion-based generator known for its enhanced prompt adherence and high-resolution output. In both cases, the resulting images lacked visual features consistent with cleft lip anatomy despite successfully generating faces. The results indicate that even advanced open-source or commercial models do not naturally associate the term "cleft lip" with the corresponding anatomical traits.

To further explore high-performance commercial systems, we evaluated FLUX 1.1 Pro, the most capable FLUX image generation suite variant. As expected, Figure 1c shows that the generated image exhibits high visual realism. However, it fails to reproduce any characteristic features of cleft lip morphology, similar to the previous models. Finally, we examined FLUX.1 [schnell], the open-source and computationally lighter variant we later use for domain-specific fine-tuning. Figure 1d shows that this model also does not generate facial features indicative of cleft lip. Furthermore, even when explicitly prompted to produce "a cleft lip scar," the model failed to associate the condition with the orofacial region.

These qualitative results highlight a critical limitation in general-purpose text-to-image models without targeted adaptation. They do not capture the medically relevant features of cleft lip, regardless of prompt specificity or model capacity.

5.2 Quantitative Metrics Analysis

Table 2 summarizes the quantitative results for the three evaluation metrics used in this study: FIR, FID, and LPIPS, computed for the generated images of both the pre-operative and post-operative cleft lip models. We use CleftGAN [11] as a baseline, generating 2,000 cleft lip images using the model released by the authors. Given that we only have the generated images and not the training images, we only estimate LPIPS metrics for these images. Unlike CleftGAN, which is limited to generating cleft lip images without control over specific attributes, our method allows prompt-based specification of cleft type, repair status, and subject's demography.

We compute FIR by measuring the cosine similarity between the embeddings of each generated image and all real training images of the corresponding class to assess potential identity leakage. We use a similarity threshold of 0.6 to classify a generated-training pair as an identity match. For the pre-operative generated images, the FIR match rate was 0.0025%, with only 5 out of 2,0000 pairs exceeding the threshold. The post-op cleft lip model performed even better, with a match rate of just 0.0002% (1 image). These extremely low match rates indicate that neither model reproduces training identities, demonstrating strong identity preservation despite the limited size of the training datasets (100 and 201 images, respectively). We visualize the full distribution of cosine similarity scores of FIR between all generated–training image pairs to investigate identity leakage further. Figure 2 shows that both distributions approximate a Gaussian curve centered at 0.18 for pre-operative and 0.10 for post-operative samples. Even with a lower threshold (e.g., 0.5), identity matches remain rare, further reinforcing the conclusion that the models do not overfit to training identities.

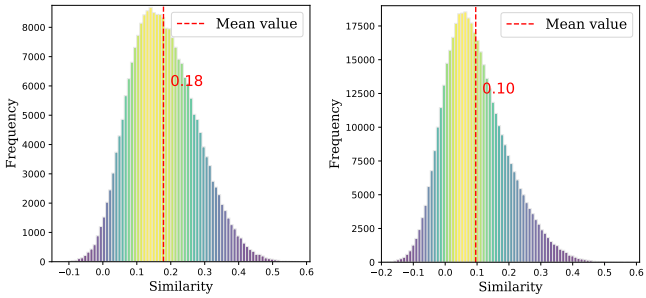
Next, we evaluate the alignment between the real and generated image distributions using FID. Lower FID values indicate higher similarity to the target distribution. Our models achieved FID scores of 7.277 for the pre-operative set, 7.774 for the post-operative set, and 5.911 considering both classes. These values indicate a high degree of visual fidelity and distributional alignment. For context, we compare our results to a recent study by Hayajneh et al. [11], which also aimed to generate cleft lip facial images. Their GAN-based approach achieved an FID of 21.03 when evaluated on 2,000 generated images compared to a training set of 514 real cleft images. While the datasets differ, our substantially lower FID values suggest that diffusion-based fine-tuning yields improved alignment with the underlying data distribution for this domain.

Finally, we computed LPIPS to evaluate perceptual diversity among the generated images. LPIPS scores range from 0 (identical images, indicating model collapse) to 1 (maximum perceptual variation). LPIPS values typically range between 0.2 and 0.6 for

Table 2: Quantitative metrics computed on 2,000 generated images per class for pre- and post-op cleft lip.

Metric	Pre-Op	Post-Op	All Cleft Lip
FIR ↓	0.0025%	0.0002%	–
FID ↓	7.277	7.774	5.911
LPIPS ↔	0.542	0.559	0.556
LPIPS: CleftGAN [11]↔	–	–	0.335

↓ indicates that the metric improves when it is lower, while ↔ shows improvement with values far from the extremes 0 or 1 (e.g., 0.5).



(a) Pre-op cleft images (b) Post-op cleft images

Figure 2: Histograms of similarity scores used to compute FIR. The show the mean value with the vertical red line.

the synthesis of faces, depending on the task constraints. Prior studies have reported reference values such as 0.42 for face generation [21], 0.39 for diffusion-based text-conditioned generation [33], and 0.2–0.4 for image editing tasks [16]. In medical imaging contexts, broader ranges such as 0.23–0.81 have been observed [37]. We aim to avoid both extremes given that our generation is prompt-conditioned and restricted to the facial domain (values too low indicative of mode collapse, and values too high suggesting excessive randomness). As reported in Table 2, our LPIPS values are 0.542 for pre-operative, 0.559 for post-operative samples, and 0.556 for both classes. These values surpass the LPICS result achieved by the CleftGAN [11] baseline, which is just 0.335. These results suggest that our approach generates sufficiently diverse yet semantically coherent outputs, capturing variation in cleft-specific anatomical features without deviating from the clinical domain.

5.3 Human Scores for Realism and Anatomical Plausibility

We conducted a human evaluation with two medical specialists experienced in treating cleft lip to assess the visual realism and anatomical plausibility of the generated images. A total of 60 images were independently rated, consisting of 50 generated images and 10 real images. Half of the images depicted pre-operative cases, and the other half post-operative cases. Each image was scored on a 5-point Likert scale (1–5), where 5 denotes the highest level of perceived realism or anatomical plausibility.

Figure 3 shows bar plots of scores for the generated images only. Post-operative generated images received particularly strong ratings, with average scores of 4.96 for realism and 4.86 for anatomical plausibility. These results confirm that the fine-tuned model successfully captures subtle anatomical outcomes of cleft lip repair,

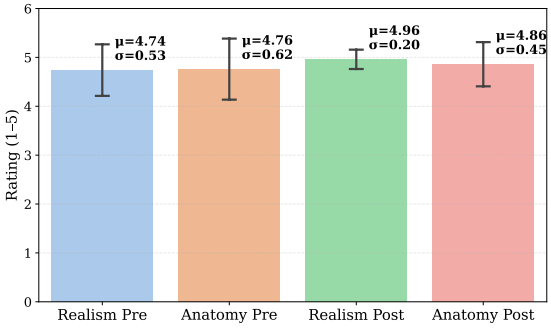


Figure 3: Human ratings for generated images, assessing visual realism and anatomical plausibility. Post-op images achieved higher average scores compared to pre-op images, indicating that the model captured subtle surgical outcomes effectively.

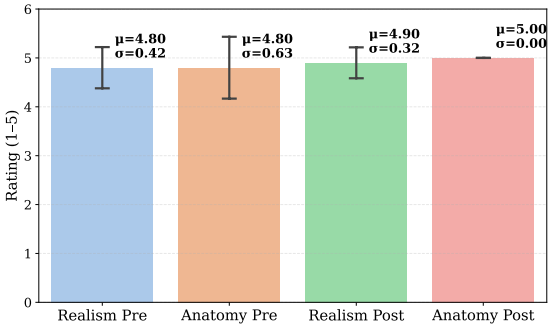
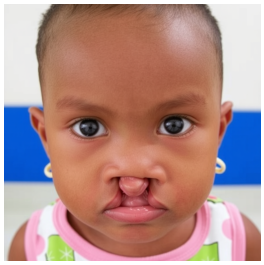


Figure 4: Human ratings for real images used in the evaluation. Pre-op real images received lower anatomical plausibility scores, supporting the hypothesis that presenting pre-op real images of older children negatively influenced expert ratings.

such as scarring and asymmetry. Both realism and anatomical plausibility for pre-operative generated images were also highly rated, with mean values above 4.7. This slight decrease in the scores for pre-operative images compared to the post-operative ones across both categories, suggests that the model produces slightly more convincing post-operative samples. One possible explanation for this discrepancy lies in the generation of children's faces, rather than exclusively baby faces, within the pre-operative group. Clinically, cleft repair is ideally performed within the first year of life, often before six months of age [9]. Consequently, the appearance of an older child with an unrepaired cleft lip may seem less anatomically plausible to medical professionals—not due to flaws in image generation, but because such presentations are uncommon in real clinical practice. This observation highlights the importance of aligning prompt design with clinical expectations, particularly in medical image synthesis tasks.

To further validate these observations, we evaluated the scores assigned to real images within the assessment set. As shown in Figure 4, the same trend persisted: pre-operative real images received lower anatomical plausibility scores. This finding strengthens our hypothesis that age-related differences, rather than the synthetic quality of the images, influenced expert ratings. Notice that the

(a) "[CLP] A kid's face with an unrepaired bilateral cleft lip"



(b) "[CLP] A girl's face with an unrepaired bilateral cleft lip. Dark skin"

(c) "[CLP] A boy's face with an unrepaired unilateral cleft lip. Light skin. Green eyes"

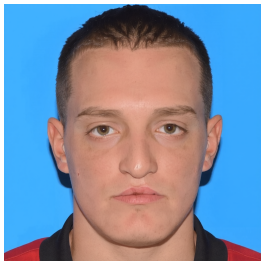
(d) "[CLP] A baby's face with an unrepaired unilateral cleft lip"

Figure 5: Examples of generated faces with a pre-operatory cleft lip.

(a) "[CLPrep] A girl's face with an unrepaired bilateral cleft lip. Dark skin"

(b) "[CLPrep] A girl's face with slight nose and lip asymmetry after cleft lip repair. Light skin"

(c) "[CLPrep] A baby's face with a scar on the philtrum after cleft lip repair"



(d) "[CLPrep] A man's face with slight nose and lip asymmetry after cleft lip repair. Short hair"

Figure 6: Examples of generated cleft lip faces showing diversity across pre- and post-op conditions.

average scores for synthetic (Figure 3) and real (Figure 4) images are very close, showing the potential of our proposed strategy.

Figures 5 and 6 show representative examples of pre- and post-operative images, alongside the prompts used to generate them. These highlight the model's ability to render text-guided facial attributes, including gender, skin tone, eye and hair color, ethnicity, age group, and cleft-related morphology. Figure 5 illustrates the generation of varying severities of bilateral cleft lip, as well as unilateral clefts with nasal asymmetry on the side of the defect, which is consistent with clinical observations. Figure 6 showcases several post-operative features: (a) nasal tip ptosis and a poorly defined Cupid's bow common among bilateral cleft repair; (b) subtle nose and lip asymmetry; (c) visible philtrum scarring; and (d) realistic adult facial structures following surgical repair. These examples demonstrate that the model can accurately synthesize both structural variation and demographic diversity in cleft lip conditions.

5.4 Embedding-Space Variance

We conduct an analysis based on *principal component analysis* (PCA) on face embeddings computed for the FIR metric to visualize the distribution of the generated facial images. Figure 8 presents the PCA visualization for real and generated faces with a cleft lip and a set of 2,000 facial images of healthy children obtained from Medvedev et al. [24]. All images were processed as described in Section 4.3.1.

Both the generated pre- and post-operative images show strong spatial alignment with their respective real counterparts. This suggests that the models have successfully learned to generate well-distributed images within the same embedding space as the real cleft lip datasets. A clear spatial distinction can be observed between pre- and post-operative faces, particularly in the horizontal (PCA 1) direction. This result indicates that the embedding model captures meaningful anatomical differences between the two classes, and that such structure is preserved in the generated data as well. Interestingly, the cluster of healthy children faces overlaps more strongly with the post-operative group than with the pre-operative one. This observation is consistent with clinical expectations, as post-op cleft lip faces tend to appear closer to normative facial morphology following surgery. This observation further validates that the generated post-operative faces are not only diverse but also converge toward realistic and healthy facial structures.

The spread of both generated pre- and post-operative faces shows a distribution comparable in scale to that of real samples, indicating that the models are not collapsing into narrow modes or repeating similar identities. This embedding-space variance supports the LPIPS findings regarding intra-class diversity.

5.5 Lip Anomaly Detection with Generated Images

To demonstrate the practical utility of the generated images, we train a CNN-based model following the single-branch architecture described in Rosero et al. [31] for lip anomaly detection. The model classifies images as either displaying an abnormal characteristic of post-op cleft lip or representing a healthy face. We train the model using generated post-op cleft lip images and healthy young face images sourced from Medvedev et al. [24]. However, we evaluate the model exclusively on a balanced set of 100 real images. This approach achieves an accuracy of 79%, substantially outperforming the 66% accuracy reported in Rosero et al. [31] synthetic data under comparable settings (i.e., typical faces with post-op cleft lip modifications).

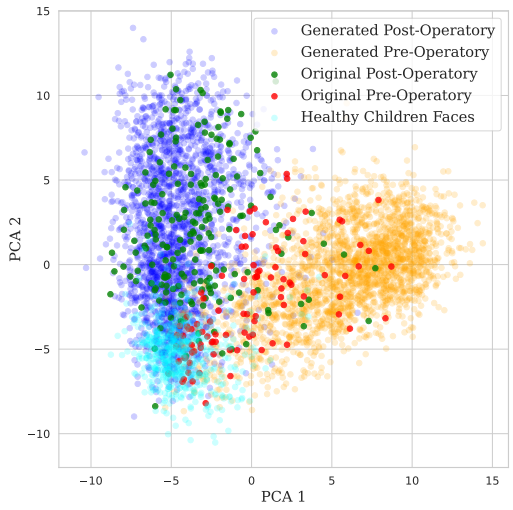
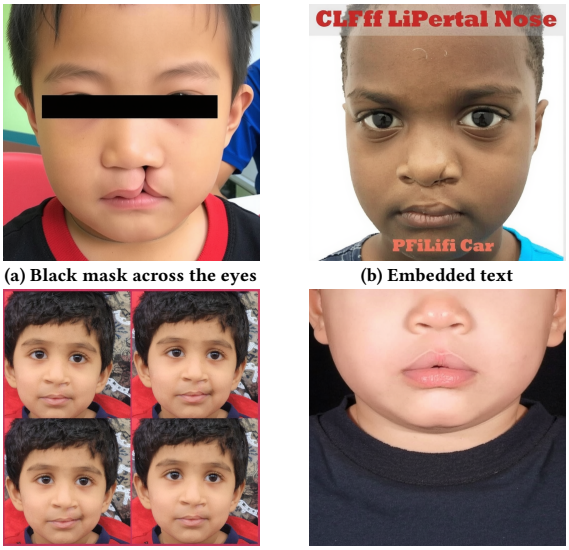


Figure 7: PCA visualization of face embeddings using the first two principal components. The figure shows real and generated pre-operatory and post-operatory images. It also shows real images from control children without a cleft lip condition.



(a) Black mask across the eyes **(b) Embedded text**

(c) Multiple faces within the same image **(d) Partially cropped facial image**

Figure 8: Limitations on the generation of pre- and post-operative images with cleft lip.

Additionally, we assess the value of the pre-op generated images by conducting a classification task between pre- and post-op cleft lip (2,000 generated images per class). We solely train on generated images. We evaluate the models on real images. We select 50 pre-op cleft lip and 50 post-op cleft lip images. The model achieves an accuracy of 84% without relying on any real samples during training. These results highlight the effectiveness of our generated datasets in supporting clinical image classification tasks and underscore their potential for future applications in data augmentation and training models where access to real clinical images is limited.

6 Limitations

While our fine-tuned models demonstrate strong performance in generating clinically plausible facial images with cleft lip features, we observed a few limitations inherent to text-to-image generation. Despite incorporating explicit instructions in each prompt, such as *"Only one face, no text, no watermarks, no masks or occlusions"*, as described in Section 3.3, a reduced set of generated images still contains undesired artifacts. These include partial occlusions (e.g., black mask across the eyes), embedded text, multiple faces within the same image frame, and partially cropped or incomplete facial regions, as illustrated in Figure 8.

These artifacts may indicate biases in the base model's pretraining data, where training images with watermarks, clinical annotations, or masked identities may be present (our training data does not present any of these artifacts). Even detailed prompts may not be enough to override these inherited priors. These limitations highlight the need for complementary post-processing pipelines for image filtering, reinforcement learning, or adversarial feedback mechanisms to penalize the generation of such artifacts further.

7 Conclusions

This work presented a method to fine-tune a state-of-the-art open-source text-to-image diffusion model to generate clinically relevant facial images depicting pre-operative and post-operative cleft lip conditions. Leveraging a small dataset of 301 labeled facial images, we employ prompt-based conditioning and LoRA to enable efficient domain-specific adaptation without the need to retrain the full model. Prompt design was carefully controlled to guide the synthesis of age, gender, and anatomy-specific features.

We evaluated identity leakage, visual fidelity, diversity, and anatomical plausibility. Quantitative results show that our method achieves low FIR match rates, indicating strong protection against training identity leakage. The generated images closely align with the distribution of real data, as reflected in low FID scores, and exhibit substantial intra-class variation according to LPIPS. Human evaluation by medical experts further confirms the high realism and anatomical accuracy of the generated samples, particularly for post-operative representations.

Our approach demonstrates its effectiveness on the downstream task of lip anomaly detection, opening opportunities for synthetic data augmentation in medical imaging domains where privacy and annotation costs still limit scalable machine learning approaches. Specifically, generated images can be used to enhance facial analysis tools, including face landmark detection, and the assessment of oro-facial symmetry after surgery, reducing the reliance on real clinical images for training and reserving them primarily for evaluation.

8 Safe and Responsible Innovation Statement

Our work aims to support medical training and diagnosis through the generation of realistic images of cleft lip conditions. We prioritize patient privacy by exclusively using synthetic data, mitigating risks associated with real facial imagery. We acknowledge potential biases in model outputs related to demographic representation and

are actively exploring ways to improve inclusivity across skin tones and facial features. While synthetic medical images offer significant benefits, we caution against misuse in contexts lacking clinical oversight. This work is intended to augment, not replace, expert decision-making and is developed with sensitivity toward ethical deployment in healthcare settings.

References

[1] Mohamed Akrout, Bálint Gyepesi, Péter Holló, Adrienn Poór, Blága Kincső, Stephen Solis, Katrina Cirone, Jeremy Kawahara, Dekker Slade, Latif Abid, et al. 2023. Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 99–109.

[2] K Anusudha et al. 2024. Real time face recognition system based on YOLO and InsightFace. *Multimedia Tools and Applications* 83, 11 (2024), 31893–31910.

[3] Daniel Anojan Atputharuban, Christoph Theopold, and Aonghus Lawlor. 2024. CleftLipGAN: Interactive GAN-Inpainting for Post-Operative Cleft Lip Reconstruction. In *Proceedings of the Asian Conference on Computer Vision*. 175–192.

[4] J.A. de Souza Freitas, L.T. das Neves, A.L.P.F. de Almeida, D.G. Garib, I.K. Trindade-Suedam, R.Y.F. Yaedú, R.d.C.M.C. Lauris, S. Soares, T.M. Oliveira, and J.H.N. Pinto. 2012. Rehabilitative treatment of cleft lip and palate: experience of the Hospital for Rehabilitation of Craniofacial Anomalies/USP (HRAC/USP)-Part 1: overall aspects. *Journal of Applied Oral Science* 20, 1 (February 2012), 9–15. doi:10.1590/S1678-77572012000100003

[5] Jiankang Deng, Jia Guo, Xiang An, Zheng Zhu, and Stefanos Zafeiriou. 2021. Masked Face Recognition Challenge: The InsightFace Track Report. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 1437–1444.

[6] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. 2020. Sub-center ArcFace: Boosting Face Recognition by Large-scale Noisy Web Faces. In *Proceedings of the IEEE Conference on European Conference on Computer Vision*.

[7] Tribikram Dhar, Nilanjan Dey, Surekha Borra, and R. Simon Sherratt. 2023. Challenges of Deep Learning in Medical Image Analysis—Improving Explainability and Trust. *IEEE Transactions on Technology and Society* 4, 1 (2023), 68–75. doi:10.1109/TTS.2023.3234203

[8] Matthew Fell, Alex Davies, Amy Davies, Shaheel Chummun, Alistair RM Cobb, Kanwalraj Moar, and Yvonne Wren. 2023. Current surgical practice for children born with a cleft lip and/or palate in the United Kingdom. *The Cleft Palate Craniofacial Journal* 60, 6 (2023), 679–688.

[9] Carrol Gamble, Christina Persson, Elisabeth Willadsen, Liz Albery, Helene Soegaard Andersen, Melissa Zattoni Antoneli, Malin Appelqvist, Ragnhild Aukner, Pia Bodling, Melanie Bowden, et al. 2023. Timing of primary surgery for cleft palate. *New England Journal of Medicine* 389, 9 (2023), 795–807.

[10] Weixiang Gao and Yifan Xia. 2024. CCFExp: Facial Image Synthesis with Cycle Cross-Fusion Diffusion Model for Facial Paralysis Individuals. *arXiv e-prints* (2024), arXiv–2409.

[11] Abdullah Hayajneh, Erchin Serpedin, Mohammad Shaqfeh, Graeme Glass, and Mitchell A Stotland. 2025. Adapting a style based generative adversarial network to create images depicting cleft lip deformity. *Scientific Reports* 15, 1 (2025), 3614.

[12] Abdullah Hayajneh, Erchin Serpedin, and Mitchell Stotland. 2024. Automatic Semantic In-Painting Image Normalization for Facial Anomaly Appraisal. In *2024 32nd European Signal Processing Conference (EUSIPCO)*. IEEE, 1501–1505.

[13] 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. *Advances in neural information processing systems* 30 (2017).

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.

[15] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-free generative adversarial networks. *Advances in neural information processing systems* 34 (2021), 852–863.

[16] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6007–6017.

[17] Firas Khader, Gustav Müller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarburger, Maximilian Schulze-Hagen, Philipp Schad, Sandy Engelhardt, Bettina Baeßler, Sebastian Foersch, et al. 2023. Denoising diffusion probabilistic models for 3D medical image generation. *Scientific Reports* 13, 1 (2023), 7303.

[18] Bardia Khosravi, Pouria Rouzrokh, John P Mickley, Shahriar Faghani, Kellen Mulford, Linjun Yang, A Noelle Larson, Benjamin M Howe, Bradley J Erickson, Michael J Taunton, et al. 2023. Few-shot biomedical image segmentation using diffusion models: beyond image generation. *Computer Methods and Programs in Biomedicine* 242 (2023), 107832.

[19] Aron Kirchhoff, Alexander Hustinx, Behnam Javanmardi, Tzung-Chien Hsieh, Fabian Brand, Fabio Hellmann, Silvan Mertes, Elisabeth André, Shahida Moosa, Thomas Schultz, et al. 2025. GestaltGAN: Synthetic photorealistic portraits of individuals with rare genetic disorders. *European Journal of Human Genetics* (2025), 1–6.

[20] Black Forest Labs. 2024. FLUX. https://github.com/black-forest-labs/flux.

[21] Yichen Li, Yufei Yin, Wengang Zhou, and Houqiang Li. 2024. Refining Video-Based Person Re-Identification: An Integrated Framework with Facial and Body Cues. In *Proceedings of the 1st ICMR Workshop on Multimedia Object Re-Identification*. 1–5.

[22] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747* (2022).

[23] C.T. Mai, J.L. Isenburg, M.A. Canfield, R.E. Meyer, A. Correa, C.J. Alverson, P.J. Lupo, T. Riehle-Colarusso, S.J. Cho, D. Aggarwal, and R.S. Kirby. 2019. National population-based estimates for major birth defects, 2010-2014. *Birth Defects Research* 111, 18 (October 2019), 1420–1435. doi:10.1002/bdr2.1589

[24] Iurii Medvedev, Farhad Shadmand, and Nuno Gonçalves. 2024. Young Labeled Faces in the Wild (YLFW): A Dataset for Children Faces Recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2024)*. IEEE, Istanbul, Turkey.

[25] Gemma S Parra-Dominguez, Raul E Sanchez-Yanez, and Carlos H Garcia-Capulin. 2021. Facial paralysis detection on images using key point analysis. *Applied Sciences* 11, 5 (2021), 2435.

[26] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4195–4205.

[27] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[29] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.

[30] K. Rosero, A. Salman, B. Sisman, R. Hallac, and C. Busso. 2024. Enhanced Facial Landmarks Detection for Patients with Repaired Cleft Lip and Palate. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2024)*. Istanbul, Turkey, 1–10. doi:10.1109/FG59268.2024.10582022

[31] Karen Rosero, Ali N Salman, Rami R Hallac, and Carlos Busso. 2024. Lip abnormality detection for patients with repaired cleft lip and palate: a lip normalization approach. In *Proceedings of the 26th International Conference on Multimodal Interaction*. 479–487.

[32] Karen Rosero, Ali N Salman, Lucas M Harrison, Alex A Kane, Carlos Busso, and Rami R Hallac. 2025. Deep Learning-Based Assessment of Lip Symmetry for Patients With Repaired Cleft Lip. *The Cleft Palate Craniofacial Journal* (2025), 10556656241312730.

[33] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 22500–22510.

[34] N. Salari, N. Darvishi, M. Heydari, S. Bokaee, F. Darvishi, and M. Mohammadi. 2022. Global prevalence of cleft palate, cleft lip and cleft palate and lip: A comprehensive systematic review and meta-analysis. *Journal of Stomatology, Oral and Maxillofacial Surgery* 123, 2 (April 2022), 110–120. doi:10.1016/j.jormas.2021.05.008

[35] Henning Schliephake and Jarg-Erich Hausamen. 2023. Cleft lip and palate. In *Oral and maxillofacial surgery: Surgical textbook and atlas*. Springer, 331–386.

[36] Sabina Umirzakova, Shabir Ahmad, Sevara Mardieva, Shakhnoza Muksimova, and Taeg Keun Whangbo. 2023. Deep learning-driven diagnosis: A multi-task approach for segmenting stroke and Bell's palsy. *Pattern Recognition* 144 (2023), 109866.

[37] Mehmet Ozan Unal, Metin Ertas, and Isa Yildirim. 2024. Proj2Proj: self-supervised low-dose CT reconstruction. *PeerJ Computer Science* 10 (2024), e1849.

[38] Tarun Vyas, Prabhakar Gupta, Sachin Kumar, Rajat Gupta, Tanu Gupta, and Harkanwal Preet Singh. 2020. Cleft of lip and palate: A review. *Journal of family medicine and primary care* 9, 6 (2020), 2621–2625.

[39] Xukang Wang, Ying Cheng Wu, Mengjie Zhou, and Hongpeng Fu. 2024. Beyond surveillance: privacy, ethics, and regulations in face recognition technology. *Frontiers in big data* 7 (2024), 1337465.

[40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

Received May 2025; revised June 2025; accepted July 2025