FaceSafe: An Inpainting Pipeline for Privacy-Compliant Scalable Image Datasets

Ananya Salian [* 1] **Sydney Su** [* 1] **Roger You** [1] **Lening Nick Cui** [1] **Patrick Cui** [1] **Charles Duong** [† 1]
Vasu Sharma [† 1] **Sean O'Brien** [† 1] **Kevin Zhu** [† 1]

Abstract

Large-scale web-scraped datasets have contributed significantly to progress in deep learning, yet the extensive presence of biometrics data, such as faces, poses a legitimate legal, ethics, and privacy issue. Existing approaches address this by removing sensitive images entirely, often sacrificing downstream performance, or purchasing use of licensed images. To address this gap, we present a novel privacy preserving transformation pipeline that uses a diffusion-based inpainting model to systematically replace detected faces in images with multiple, synthetic variants conditioned on different demographic attributes, resulting in a novel, privacy-preserving dataset of distinct face images. Our method, evaluated on $12,000$ images transformed from LAION-400M and CelebA-HQ, eliminates privacy risks without significant loss of image quality or diversity. This transformation pipeline will serve as a scalable guideline for the creation of datasets that follow legal and ethical privacy constraints.

1. Introduction

The creation of nonconsensually collected image datasets in recent years [1, 2, 3, 4, 5, 6] has undoubtedly spurred progress in computer vision and generative modeling [7]. Furthermore, datasets specifically created through web-scraping can sometimes provide up to billions of image-to-text pairs for training state-of-the-art models [6]. However, the inclusion of biometric identifiers, such as human faces, poses significant ethical, privacy, and legal challenges, especially given recent legislative trends in the United States towards stricter biometric data regulation [8]. Recent privacy laws introduced and enacted in several U.S. states impose strict requirements on the collection, processing, and retention of biometric data, and mandate explicit consent for collection, processing, and retention of these images [9]. Specifically, recent privacy laws such as the Illinois Biometric Information Privacy Act (BIPA)[1] and the Texas Biometric Identifier Act[2] are some of the most restrictive, completely barring any usage of biometrics data compiled nonconsensually, and fining up to $\$5,000$ at minimum per infraction. These laws render most of the aforementioned datasets unusable on a legally compliant basis, unless all subjects are explicitly determined to not be residents of Illinois, Texas, or other states with similar laws, and thus provide a legal motivation for the development of methods that can transform an existing dataset to be legally compliant.

Existing approaches in privacy efforts primarily remove, filter, or label certain images, and can be automated either algorithmically or through the use of AI agents [10, 11, 7]. Other types of dataset curation, such as annotating images, can be similarly done algorithmically [12], or done through manual, crowd-sourced labor [13, 14, 15]. While effective in certain use cases such as the ethical concern of including children in datasets, these approaches generally risk excessive data loss, may fail to address more nuanced privacy concerns, and in the case of manual human curation, may be too expensive for large datasets [7]. Therefore, this calls for a more robust method that can properly confront legal privacy concerns and dataset functionality while maintaining relatively low-costs at a large-scale.

To address this challenge, we propose a novel compliance-by-transformation pipeline tailored to any dataset containing nonconsensually imaged faces. We demonstrate our pipeline on subsets of the LAION-400M and CelebA-HQ datasets [6, 16], and provide further analyses and case studies on our produced images.

From a high-level overview, our pipeline first uses a pre-trained Multi-task Cascaded Convolutional Network (MTCNN) [17] to detect the probability of a face in a picture and filter out images where the resolution of the face is too small. Then, a blank elliptical mask is overlayed on top of the detected face region, and a stable diffusion model [18] finetuned on the FairFace dataset [19] is used to in-

[*] Equal contribution † Equal Senior Author [1] Algoverse AI Research. Correspondence to: Ananya Salian <ananya.salian@student.unimelb.edu.au>.

Published at Data in Generative Models Workshop: The Bad, the Ugly, and the Greats (DIG-BUGS) at ICML 2025, Vancouver, Canada. Copyright 2025 by the author(s).

[1] 740 Ill. Comp. Stat. Ann. 14/15

[2] Tex. Bus. & Com. Code Ann. §503.001 et seq.

paint the region. Each face is then inpainted with 8 distinct replacements via conditioned prompts that combine the estimated age group, gender, and ethnicity. These prompts were empirically determined to be the most effective as they had a high Fréchet Inception Distance (FID) [20] score against each other. If multiple faces are in an image, then additional permutations are generated, and the best 8 overall images are selected. This style of combinatorial inpainting encourages the final dataset to be more diverse, and ultimately each original face image is replaced by multiple legally compliant augmented variants, solving the issue of legal compliance.

Finally, to assess the quality of our inpainted outputs, we conduct a multi-metric evaluation focused on both identity anonymization and semantic preservation. We use ArcFace [21] to quantify identity similarity between original and inpainted images, PSNR [22], an image quality metric that identifies reconstruction error, CLIP similarity, [23], which evaluates how edited images retain their prompting attributes, LPIPS [24], to assess image realism and quality, FID [20], to benchmark the overall quality of our inpainted samples against real images, and SSIM/MS-SSIM [25, 26, 24], to evaluate structural similarity at multiple scales. We also assess demographic consistency before and after inpainting to ensure minimal changes in race, gender, and age distribution.

Through this pipeline, we demonstrate that a large non-compliant image dataset can be successfully and practically transformed to be legally compliant. Our work provides both a practical methodology and a reusable benchmark for those striving to utilize privacy compliant data.

2. Related Work

2.1. Diffusion Models

Diffusion models are a powerful class of generative models that have significantly advance the state-of-the-art across diverse domains, particularly in image synthesis, super-resolution, inpainting, and semantic editing [27, 28]. There are several approaches for creating diffusion models [29, 30, 31, 27], however, most notably, our inpainting pipeline primarily utilizes Denoising Diffusion Probabilistic Models (DDPMs), which define a fixed forward noising schedule over T steps and learn a reverse Markovian denoising process.

Inpainting with diffusion models leverages the same generative principles but also introduces spatial awareness when reconstructing masked regions. Initial attempts focused on texture propagation [32], while more recent methods like Region-Aware Diffusion (RAD) [33] introduce mask-sensitive noise scheduling for better spatial control.

Additionally, to improve efficiency and controllability, Denoising Diffusion Implicit Models (DDIM) [34] reinterprets the reverse process as a non-Markovian trajectory, enabling 10x-50x fewer sampling steps with minimal fidelity loss, and uses stochastic differential equation (SDE) solvers to further reduce inference latency [31]. In addition, by shifting diffusion into a learned latent space, models incur 5×–10× lower compute and memory costs for high-resolution synthesis, as demonstrated by Latent Diffusion Models [18].

Finally, large-scale implementations such as Flux [35] and Stable Diffusion XL (SDXL) [36] demonstrate the scalability and generalizability of latent diffusion, producing high-resolution imagery with strong semantic fidelity.

2.2. Inpainting

Image inpainting refers to the task of reconstructing missing or occluded regions within an image by leveraging contextual information from the surrounding areas. Inpainting has evolved from classical patch-based algorithms to modern learning-based systems [37, 32].

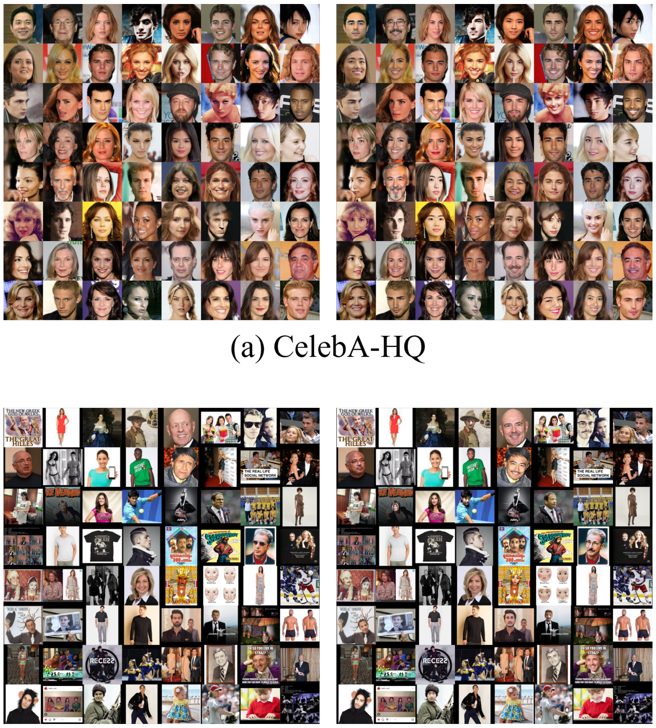
Traditional inpainting techniques—such as diffusion-often fail to capture high-level semantics, particularly in structured regions like faces. However, advances in deep learning have overcome this by modeling global context and learning structural priors from large-scale datasets. Early diffusion-based models [38] introduced denoising diffusion probabilistic models (DDPMs) as a generative framework that iteratively refines noisy inputs into realistic images. Further improvements such as contextual attention mechanisms [39]demonstrated the benefits of guiding inpainting with semantically aligned context.

RAD [33] (Region-Aware Diffusion) extends DDPMs by incorporating spatially variant noise schedules aligned with the structure of the inpainting mask, allowing pixel-wise control during denoising. To maintain realism and semantic consistency in face inpainting, Generative Facial Prior for Blind Face Restoration [40] (GPEN) leverages generative facial priors—such as identity embeddings, landmarks, or attribute-aware constraints—to guide facial reconstruction. While effective for enhancing visual fidelity, such methods often seek to preserve or restore identifiable facial features, making them less suitable for privacy-critical applications.

2.3. Existing Privacy Methods

A growing body of work explores generative techniques for preserving privacy, particularly in the context of facial data [10, 41, 42, 43]. One line of work uses crowd-sourced annotations to apply obfuscation masks to faces, demonstrating that model utility can be preserved even with significant visual distortion. However, contrary to our focus, this study utilized manual labor, which scales impractically [10]. These masks also typically obscure rather than replace identity-bearing features, maximally limiting downstream utility for certain tasks, such as training generative models. More recent efforts such as Diff-Privacy [41] utilize pixel-level operations and multi-scale inversion modules that iteratively distort facial features. These transformations achieve high visual fidelity while intentionally rendering the identity unrecognizable to both human observers and machine learning models. Another method, DIFP [43], additionally builds on this idea by introducing a conditionally guided face generator that produces encrypted facial images. Its two-stage pipeline can realistically regenerate facial features while maintaining the ability to decrypt the original identity using a reverse diffusion process. This reversible nature, however, limits its applicability under strict privacy regulations that mandate irreversibility. ID^3 [42] takes a different approach, focusing on generating synthetic but identity-consistent faces to augment recognition datasets, primarily focusing on diversity and realism over anonymization.

3. Inpainting Methodology: FaceSafe


(a) CelebA-HQ

(b) LAION

Figure 1: Snapshots of original images on the left and their corresponding anonymized transformations on the right for CelebA-HQ and LAION-400M.

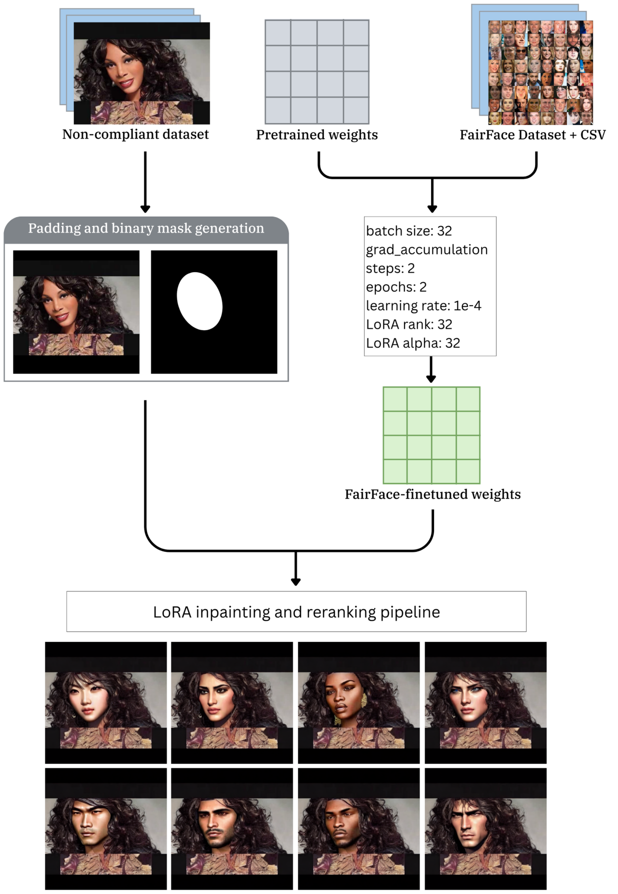


Figure 2: A simplified flow diagram of our inpainting transformation pipeline.

To enable the construction of large-scale, privacy-compliant datasets, we introduce FaceSafe, an inpainting pipeline designed for scalable face anonymization. This section outlines the design and implementation of FaceSafe, shown in Fig. 2. Notable components include training of a demographically-conditioned LoRA-injected Stable Diffusion model, the preprocessing steps for face detection and mask creation, and the inference-time re-ranking strategy that ensures both visual realism and semantic fidelity. In addition, we provide information on the scalability of the system and analyze the demographic fidelity and obfuscation of the identity of the resulting dataset.

To build a demographically diverse inpainting model, we begin by curating 44k portraits from the FairFace dataset [19] spanning four coarse race labels (Asian, Caucasian, African, Middle Eastern) and two binary genders. Each combination is mapped to an eight-way one-hot vector used to condition a LoRA-injected Stable Diffusion UNet [44].

We then initialize from the 512×512 Stable Diffusion Inpainting model [18], freezing the VAE and text encoder. LoRA adapters (rank=32, $\alpha = 32$) are inserted into all nn. Linear and 1×1 nn. Conv2D layers in the UNet, resulting in around 7M trainable parameters. The model is fine-tuned for two epochs on a single A100 (40GB, FP16), using AdamW [45] ($\beta = 0.9/0.999$, learning rate = 1e-4) with cosine decay and linear warm-up over the first 5% of 2080 optimizer steps. A total batch size of 64 is achieved via gradient accumulation (32 physical × 2).

At inference time, we apply MTCNN [17] (confidence ≥ 0.9, area $\geq 0.1\%$) to detect faces, expanding each region with a 1.3× elliptical mask. For the largest detected face in each image, we generate eight demographic prompts (race × gender), sampling each four times with DPM-Solver [46] (40 Karras steps, guidance scale = 7.5, FP16, fixed seed), yielding 32 candidates. For all remaining faces, two demographic prompts (randomized race x gender) are sampled two times each with the same parameters. DeepFace [47] provides an age description token (baby, child, teenager, middle-aged adult, senior adult), and a Mediapipe [48] yaw suffix (“facing left/right”) is appended when $|\text{yaw}| \geq 15°$. Each candidate face crop is scored by four metrics:

CLIP similarity (ViT-L/14) [23] determines the semantic difference between two pieces of content by jointly embedding images and texts into the same vector space. Vectors are then compared using cosine similarity.

LPIPS-VGG (flipped to similarity) [24], a learned metric designed to measure the perceptual difference between images by comparing them in the deep feature space of pretrained convolutional neural networks.

SSIM [25, 24], a perceptual metric that measures the similarity between two images by comparing structural information, luminance, and contrast across local patches. The equation is not described for the sake of brevity, but can be found in [25].

Landmark overlap (Mediapipe) [48], which measures the degree of overlap between facial landmarks predicted by Mediapipe on both the original and inpainted face crop. We used weights of 0.4 / 0.2 / 0.2 / 0.2 for specific landmark groups, where high overlap implies spatial configuration is maintained, even if identity has been changed.

Scores are min-max normalized per face, missing values are skipped, and remaining weights are re-normalized. The top candidate for each demographic prompt is then pasted onto the image. If the masked crop contains $\geq 2\%$ high-brightness pixels (value ¿ 240), the next three best seeds are sampled to avoid “shininess” artifacts. For images with multiple faces, all permutations of top candidates are composed, and the 16 highest-scoring full-image composites are retained.

3.1. Scaling Study

To assess the feasibility of deploying our inpainting pipeline on large-scale datasets, we conducted a performance analysis using an NVIDIA A100 GPU (40GB). On average, our diffusion-based face inpainting system processes approximately 1 face every 3 seconds without candidate reranking. This timing assumes default parameters (40 inference steps, guidance scale of 7.5) and a batch size of 40.

At this throughput, the system is capable of inpainting approximately 28,800 faces per day per GPU, or roughly 1 million faces per 35 GPU-days. For context, scaling this to a web-scale dataset such as LAION-400M, assuming approximately 5% of images contain detectable human faces, would result in around 20 million candidate images. Inpainting this volume would require roughly 2,100 GPU-days, or under 3 weeks on a 100-GPU cluster, assuming full parallelization and minimal I/O bottlenecks.

We also evaluated how architectural and algorithmic choices affect scalability. One key tradeoff arises from candidate reranking, which compares multiple generated outputs per face using perceptual and semantic similarity metrics (e.g., LPIPS, SSIM, CLIP). While reranking improves visual quality and prompt alignment, it increases per-face runtime by a factor of 2–3× depending on the number of candidates and scoring metrics enabled. Therefore, running the model with reranking disabled and a single deterministic generation path is more viable for ultra-scale anonymization tasks.

Overall, the modular design of our pipeline—supporting batched inference, optional reranking, and prompt-conditioned synthesis—makes it adaptable to a wide range of scales and privacy requirements. These findings suggest that ethically-aligned, privacy-preserving data transformations on modern web-scale datasets are computationally feasible using existing diffusion-based infrastructure.

3.2. Dataset Analysis

To assess the quality of our inpainted outputs for our subsets of both LAION-400M and CelebA-HQ, we evaluate using ArcFace, PSNR, CLIP, LPIPS, FID, SSIM, and MS-SSIM. ArcFace and FID assess privacy; PSNR, SSIM, and LPIPS measure visual and pixel similarity.

We analyzed 1,000 LAION images containing 1,686 faces to evaluate the effectiveness of our face inpainting pipeline for privacy preservation. Across similarity metrics, the results indicate strong identity anonymization while maintaining reasonable perceptual and semantic similarity.

For LAION, ArcFace similarity averaged 0.16, confirming that inpainted faces were largely unidentifiable compared to their originals. CLIP similarity remained moderately high at 0.74, suggesting that semantic features such as expression, pose, or clothing were preserved. LPIPS (0.34) and PSNR (14.74 dB) confirmed perceptual divergence without severe degradation. Structural similarity metrics further supported this: SSIM averaged 0.40 and MS-SSIM 0.42, reflecting moderate but visible changes in face structure after inpainting. The FID score for LAION was 25.99, indicating moderate distributional divergence between inpainted and original images—consistent with real-world variation in lighting, orientation, and background.

Table 1: LAION: Demographic Distribution and Metrics

Demographic	**Original**	**Inpainted**
Race		
White	1251	1187
Latino/Hispanic	146	239
Black	169	154
East Asian	120	106
Gender		
Male	971	985
Female	715	701
Similarity Metric	**Mean**	**Q1–Q3**
ArcFace	0.161	0.056–0.221
PSNR (dB)	14.74	12.64–16.82
CLIP	0.739	0.677–0.813
LPIPS	0.340	0.255–0.404
SSIM	0.405	0.132–0.672
MS-SSIM	0.421	0.204–0.628
Match Rate (%)		
Age (Exact)	6.11	
Age (±5 yrs)	46.03	
Gender	67.38	
Race	68.98	
		FID: 25.99

Demographic match rates on LAION showed moderate preservation: race matched in 69%, gender in 67%, and age within ±5 years in 46% of cases. However, we observed demographic drift: a notable increase in Latino/Hispanic representation (+93) and a small increase in male representation suggest the inpainting model may inject subtle bias under unconstrained conditions.

To contextualize these findings, we also evaluated CelebA, a dataset with well-lit, front-facing portraits and centered, large faces, which presents a simpler inpainting scenario. On 1,000 CelebA images, ArcFace similarity remained similar (0.16), indicating comparable identity removal. However, all other metrics improved: PSNR increased to 21.59 dB, LPIPS dropped to 0.07, and SSIM/MS-SSIM rose to 0.82, indicating that inpainted faces in CelebA were far more visually and structurally similar to the originals. The FID for CelebA was slightly lower at 17.55, reflecting a close alignment in feature space between the inpainted and original distributions due to the dataset's visual consistency.

Table 2: CelebA-HQ: Demographic Distribution and Metrics

Demographic	**Original**	**Inpainted**
Race		
White	688	414
Latino/Hispanic	142	386
Black	97	140
East Asian	73	60
Gender		
Male	435	455
Female	565	545
Similarity Metric	**Mean**	**Q1–Q3**
ArcFace	0.160	0.095–0.216
PSNR (dB)	21.59	20.12–23.06
CLIP	0.702	0.633–0.778
LPIPS	0.066	0.049–0.078
SSIM	0.824	0.800–0.853
MS-SSIM	0.821	0.786–0.872
Match Rate (%)		
Age (Exact)	5.50	
Age (±5 yrs)	53.40	
Gender	76.20	
Race	55.80	
		FID: 17.55

Demographic match rates were higher on CelebA: gender accuracy rose to 76%, loose age match to 53%, and race match to 56%. Notably, there was no major demographic drift – the balance of race and gender remained stable between the original and inpainted sets.

These results highlight an important distinction: inpainting performance and demographic stability are highly dataset-dependent. While LAION's diverse, uncontrolled imagery tests robustness, CelebA offers a more forgiving benchmark with simpler geometry. Our approach achieves strong anonymization in both cases, but further tuning—especially on real-world data like LAION—is needed to reduce identity leakage and demographic bias under challenging conditions.

3.2.1. Case Studies

We further perform two case studies on our dataset, exploring the strengths (shown in Table 3 and limitations (shown in Fig. 3) of our inpainting pipeline.

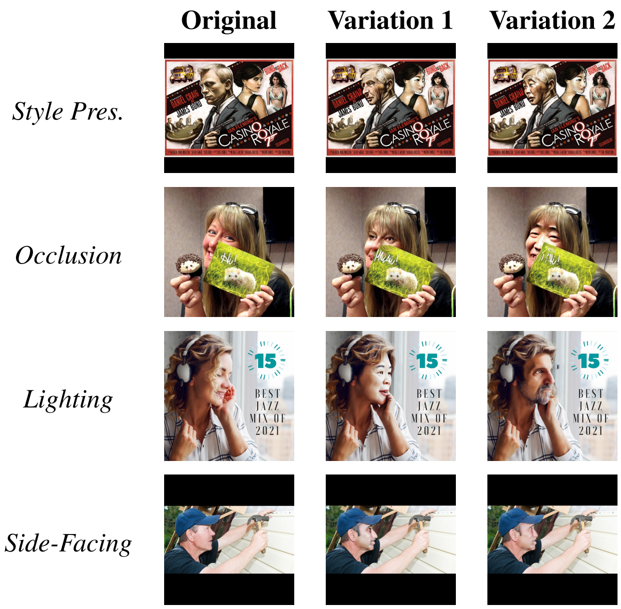


Table 3: Comparison across four facial variations: style preservation, occlusion handling, lighting variation, and side-facing pose. These results suggest that our model is effective at handling diverse real-world scenarios.

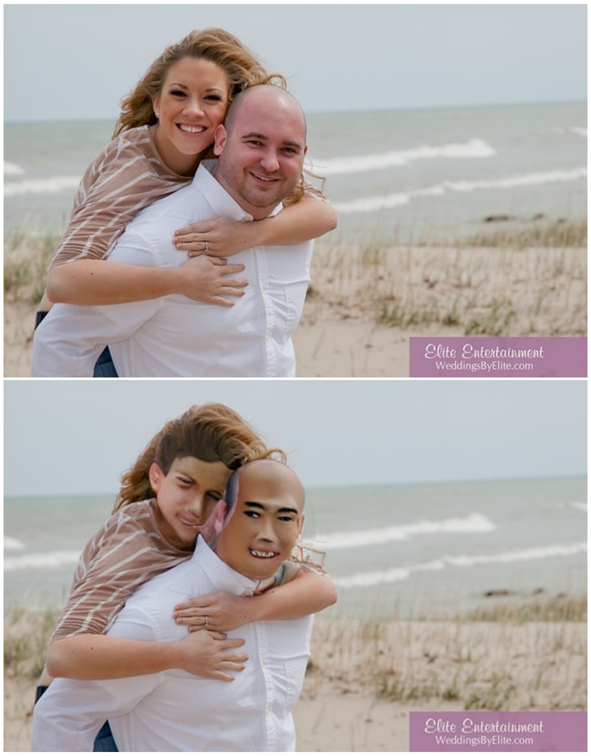


Figure 3: An example of a failed case. In close proximity, as seen here, facial overlap led to identity blending.

4. Discussion and Future Work

Despite our work's potential in practical utility, as well as demonstrated performance, our approach has several limitations that we address in this section.

Our inpainting prompts are limited to broad demographic attributes, which do not entirely capture the features of each individual. As a result, it may reinforce stereotypes (Tables 1-2), particularly with low-quality images where the model struggles to infer nuances (Fig. 1b). We plan to mitigate this by generating more detailed captions and anticipate that future diffusion models will offer improved inpainting.

Additionally, while our method seeks to anonymize faces by replacing them entirely, we do not formally verify the absence of remaining biometric features that may overlap with the original subject. In the current stage of this work, we cannot guarantee provable non-reidentifiability in all cases.

There are also additional computation and throughput constraints. Our full pipeline, while manageable for inpainting several thousand images, will currently require substantial GPU resources to process massive datasets like LAION-400M. However, our current inpainting pipeline is not fully optimized, inpainting images sequentially instead of in batches. We plan on implementing this to further streamline the process.

The pipeline also relies on several external tools for face detection, age estimation, and pose detection. These tools can introduce inaccuracies that may propagate downstream, affecting prompt quality or causing dropped samples due to failed detections.

Furthermore, although reranking reduces artifacts such as overexposure, misalignment, or complete failure, these unsuccessful image still exist, particularly in complex lighting conditions or occluded faces. These artifacts may reduce image realism or inject noise into downstream tasks unless filtered carefully. To address this, we plan on adding an additional FID filter and systematically removing images that are deemed to be unrealistic.

5. Conclusion

This paper introduces a scalable, practical data transformation pipeline, using diffusion and inpainting models, for converting non-consensually collected images into a privacy compliant dataset, and evaluates the quality of 12,000 generated images. Our work provides a blueprint for a practical dataset transformation pipeline aligned with emerging privacy legislation, offering a plan of action for ethically and legally training AI models at scale.

References

[1] Mislav Grgic, Kresimir Delac, and Sonja Grgic. “SCface – surveillance cameras face database”. en. In: *Multimedia Tools and Applications* 51.3 (Feb. 2011), pp. 863–879. ISSN: 1573-7721. DOI: 10.1007/s11042-009-0417-2. URL: https://doi.org/10.1007/s11042-009-0417-2 (visited on 05/15/2025).

[2] Manuel Günther et al. “Unconstrained Face Detection and Open-Set Face Recognition Challenge”. In: *2017 IEEE International Joint Conference on Biometrics (IJCB)*. arXiv:1708.02337 [cs]. Oct. 2017, pp. 697–706. DOI: 10.1109/BTAS.2017.8272759. URL: http://arxiv.org/abs/1708.02337 (visited on 05/15/2025).

[3] Inioluwa Deborah Raji and Genevieve Fried. *About Face: A Survey of Facial Recognition Evaluation*. arXiv:2102.00813 [cs]. Feb. 2021. DOI: 10.48550/arXiv.2102.00813. URL: http://arxiv.org/abs/2102.00813 (visited on 05/15/2025).

[4] Ergys Ristani et al. *Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking*. arXiv:1609.01775 [cs]. Sept. 2016. DOI: 10.48550/arXiv.1609.01775. URL: http://arxiv.org/abs/1609.01775 (visited on 05/15/2025).

[5] Christoph Schuhmann et al. *LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs*. arXiv:2111.02114 [cs]. Nov. 2021. DOI: 10.48550/arXiv.2111.02114. URL: http://arxiv.org/abs/2111.02114 (visited on 05/15/2025).

[6] Christoph Schuhmann et al. *LAION-5B: An open large-scale dataset for training next generation image-text models*. arXiv:2210.08402 [cs]. Oct. 2022. DOI: 10.48550/arXiv.2210.08402. URL: http://arxiv.org/abs/2210.08402 (visited on 05/10/2025).

[7] Jerone Andrews et al. “Ethical Considerations for Responsible Data Curation”. en. In: Nov. 2023. URL: https://openreview.net/forum?id=Qf8uzIT1OK (visited on 05/15/2025).

[8] Jane Horvath et al. *U.S. Cybersecurity and Data Privacy Review and Outlook – 2025*. Tech. rep. Gibson Dunn, Mar. 2025. URL: https://www.gibsondunn.com/us-cybersecurity-and-data-privacy-review-and-outlook-2025/ (visited on 05/09/2025).

[9] Woodrow Hartzog. *BIPA: The Most Important Biometric Privacy Law in the US?* en. SSRN Scholarly Paper. Rochester, NY, Oct. 2020. URL: https://papers.ssrn.com/abstract=3722053 (visited on 05/15/2025).

[10] Kaiyu Yang et al. *A Study of Face Obfuscation in ImageNet*. June 9, 2022. DOI: 10.48550/arXiv.2103.06191. arXiv: 2103.06191 [cs]. URL: http://arxiv.org/abs/2103.06191 (visited on 05/16/2025). Pre-published.

[11] Carlos Caetano et al. *Neglected Risks: The Disturbing Reality of Children’s Images in Datasets and the Urgent Call for Accountability*. arXiv:2504.14446 [cs] version: 1. Apr. 2025. DOI: 10.48550/arXiv.2504.14446. URL: http://arxiv.org/abs/2504.14446 (visited on 05/10/2025).

[12] Wenyan Li et al. *The Role of Data Curation in Image Captioning*. arXiv:2305.03610 [cs]. Feb. 2024. DOI: 10.48550/arXiv.2305.03610. URL: http://arxiv.org/abs/2305.03610 (visited on 05/10/2025).

[13] Kevin Crowston. “Amazon Mechanical Turk: A Research Tool for Organizations and Information Systems Scholars”. en. In: *Shaping the Future of ICT Research. Methods and Approaches*. Ed. by Anol Bhattacherjee and Brian Fitzgerald. Berlin, Heidelberg: Springer, 2012, pp. 210–221. ISBN: 978-3-642-35142-6. DOI: 10.1007/978-3-642-35142-6_14.

[14] Jia Deng et al. “Scalable multi-label annotation”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’14. New York, NY, USA: Association for Computing Machinery, Apr. 2014, pp. 3099–3102. ISBN: 978-1-4503-2473-1. DOI: 10.1145/2556288.2557011. URL: https://doi.org/10.1145/2556288.2557011 (visited on 05/10/2025).

[15] Zeyad Emam et al. *On The State of Data In Computer Vision: Human Annotations Remain Indispensable for Developing Deep Learning Models*. arXiv:2108.00114 [cs]. July 2021. DOI: 10.48550/arXiv.2108.00114. URL: http://arxiv.org/abs/2108.00114 (visited on 05/10/2025).

[16] Tero Karras et al. *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. arXiv:1710.10196 [cs]. Feb. 2018. DOI: 10.48550/arXiv.1710.10196. URL: http://arxiv.org/abs/1710.10196 (visited on 05/15/2025).

[17] Kaipeng Zhang et al. “Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks”. In: *IEEE Signal Processing Letters* 23.10 (Oct. 2016). arXiv:1604.02878 [cs], pp. 1499–1503. ISSN: 1070-9908, 1558-2361. DOI: 10.1109/LSP.2016.2603342. URL: http://arxiv.org/abs/1604.02878 (visited on 05/05/2025).

[18] Robin Rombach et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. arXiv:2112.10752 [cs]. Apr. 2022. DOI: 10.48550/arXiv.2112.10752. URL: http://arxiv.org/abs/2112.10752 (visited on 05/10/2025).

[19] Kimmo Kärkkäinen and Jungseock Joo. *FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age*. arXiv:1908.04913 [cs]. Aug. 2019. DOI: 10.48550/arXiv.1908.04913. URL: http://arxiv.org/abs/1908.04913 (visited on 05/10/2025).

[20] Martin Heusel et al. *GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium*. arXiv:1706.08500 [cs]. Jan. 2018. DOI: 10.48550/arXiv.1706.08500. URL: http://arxiv.org/abs/1706.08500 (visited on 05/10/2025).

[21] Jiankang Deng et al. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.10 (Oct. 2022). arXiv:1801.07698 [cs], pp. 5962–5979. ISSN: 0162-8828, 2160-9292, 1939-3539. DOI: 10.1109/TPAMI.2021.3087709. URL: http://arxiv.org/abs/1801.07698 (visited on 05/16/2025).

[22] Alain Horé and Djemel Ziou. "Image Quality Metrics: PSNR vs. SSIM". In: *2010 20th International Conference on Pattern Recognition*. ISSN: 1051-4651. Aug. 2010, pp. 2366–2369. DOI: 10.1109/ICPR.2010.579. URL: https://ieeexplore.ieee.org/document/5596999 (visited on 05/16/2025).

[23] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. Feb. 26, 2021. DOI: 10.48550/arXiv.2103.00020. arXiv: 2103.00020 [cs]. URL: http://arxiv.org/abs/2103.00020 (visited on 05/16/2025). Pre-published.

[24] Richard Zhang et al. *The Unreasonable Effectiveness of Deep Features as a Perceptual Metric*. Apr. 10, 2018. DOI: 10.48550/arXiv.1801.03924. arXiv: 1801.03924 [cs]. URL: http://arxiv.org/abs/1801.03924 (visited on 05/16/2025). Pre-published.

[25] Zhou Wang et al. "Image Quality Assessment: From Error Visibility to Structural Similarity". In: *IEEE Transactions on Image Processing* 13.4 (Apr. 2004), pp. 600–612. ISSN: 1941-0042. DOI: 10.1109/TIP.2003.819861. URL: https://ieeexplore.ieee.org/document/1284395 (visited on 05/16/2025).

[26] Z. Wang, E.P. Simoncelli, and A.C. Bovik. "Multiscale structural similarity for image quality assessment". In: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Vol. 2. Nov. 2003, 1398–1402 Vol.2. DOI: 10.1109/ACSSC.2003.1292216. URL: https://ieeexplore.ieee.org/document/1292216 (visited on 05/16/2025).

[27] Yang Song et al. "Score-Based Generative Modeling through Stochastic Differential Equations". In: *ArXiv* abs/2011.13456 (2020). URL: https://api.semanticscholar.org/CorpusID:227209335.

[28] Ling Yang et al. *Diffusion Models: A Comprehensive Survey of Methods and Applications*. arXiv:2209.00796 [cs]. Dec. 2024. DOI: 10.48550/arXiv.2209.00796. URL: http://arxiv.org/abs/2209.00796 (visited on 05/15/2025).

[29] Florinel-Alin Croitoru et al. "Diffusion Models in Vision: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.9 (Sept. 2023). arXiv:2209.04747 [cs], pp. 10850–10869. ISSN: 0162-8828, 2160-9292, 1939-3539. DOI: 10.1109/TPAMI.2023.3261988. URL: http://arxiv.org/abs/2209.04747 (visited on 05/15/2025).

[30] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851. URL: https://proceedings.neurips.cc/paper_files/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html (visited on 05/15/2025).

[31] Yang Song and Stefano Ermon. *Generative Modeling by Estimating Gradients of the Data Distribution*. Oct. 10, 2020. DOI: 10.48550/arXiv.1907.05600. arXiv: 1907.05600 [cs]. URL: http://arxiv.org/abs/1907.05600 (visited on 05/15/2025). Pre-published.

[32] Christine Guillemot and Olivier Le Meur. "Image Inpainting : Overview and Recent Advances". In: *IEEE Signal Processing Magazine* 31.1 (Jan. 2014), pp. 127–144. ISSN: 1558-0792. DOI: 10.1109/MSP.2013.2273004. URL: https://ieeexplore.ieee.org/document/6678248 (visited on 05/16/2025).

[33] Sora Kim, Sungho Suh, and Minsik Lee. *RAD: Region-Aware Diffusion Models for Image Inpainting*. Dec. 19, 2024. DOI: 10.48550/arXiv.2412.09191. arXiv: 2412.09191 [cs]. URL: http://arxiv.org/abs/2412.09191 (visited on 05/16/2025). Pre-published.

[34] Jiaming Song, Chenlin Meng, and Stefano Ermon. *Denoising Diffusion Implicit Models*. Oct. 5, 2022. DOI: 10.48550/arXiv.2010.02502. arXiv: 2010.02502 [cs]. URL: http://arxiv.org/abs/2010.02502 (visited on 05/15/2025). Pre-published.

[35] Chenglin Yang et al. *1.58-Bit FLUX*. Dec. 24, 2024. DOI: 10.48550/arXiv.2412.18653. arXiv: 2412.18653 [cs]. URL: http://arxiv.org/abs/2412.18653 (visited on 05/15/2025). Pre-published.

[36] Patrick Esser et al. *Scaling Rectified Flow Transformers for High-Resolution Image Synthesis*. Mar. 5, 2024. DOI: 10.48550/arXiv.2403.03206. arXiv: 2403.03206 [cs]. URL: http://arxiv.org/abs/2403.03206 (visited on 05/15/2025). Pre-published.

[37] Weize Quan et al. *Deep Learning-based Image and Video Inpainting: A Survey*. arXiv:2401.03395 [cs]. Jan. 2024. DOI: 10.48550/arXiv.2401.03395. URL: http://arxiv.org/abs/2401.03395 (visited on 05/15/2025).

[38] Jascha Sohl-Dickstein et al. *Deep Unsupervised Learning Using Nonequilibrium Thermodynamics*. Nov. 18, 2015. DOI: 10.48550/arXiv.1503.03585. arXiv: 1503.03585 [cs]. URL: http://arxiv.org/abs/1503.03585 (visited on 05/16/2025). Pre-published.

[39] Jiahui Yu et al. *Generative Image Inpainting with Contextual Attention*. Mar. 21, 2018. DOI: 10.48550/arXiv.1801.07892. arXiv: 1801.07892 [cs]. URL: http://arxiv.org/abs/1801.07892 (visited on 05/16/2025). Pre-published.

[40] Songhua Liu et al. *Paint Transformer: Feed Forward Neural Painting with Stroke Prediction*. Aug. 11, 2021. DOI: 10.48550/arXiv.2108.03798. arXiv: 2108.03798 [cs]. URL: http://arxiv.org/abs/2108.03798 (visited on 05/16/2025). Pre-published.

[41] Xiao He et al. *Diff-Privacy: Diffusion-based Face Privacy Protection*. Sept. 11, 2023. DOI: 10.48550/arXiv.2309.05330. arXiv: 2309.05330 [cs]. URL: http://arxiv.org/abs/2309.05330 (visited on 05/15/2025). Pre-published.

[42] Shen Li et al. *ID^3: Identity-Preserving-yet-Diversified Diffusion Models for Synthetic Face Recognition*. Version 1. Sept. 26, 2024. DOI: 10.48550/arXiv.2409.17576. arXiv: 2409.17576 [cs]. URL: http://arxiv.org/abs/2409.17576 (visited on 05/15/2025). Pre-published.

[43] Xingyi You et al. "Generation of Face Privacy-Protected Images Based on the Diffusion Model". In: *Entropy* 26.6 (May 31, 2024), p. 479. ISSN: 1099-4300. DOI: 10.3390/e26060479. PMID: 38920488. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11202580/ (visited on 05/15/2025).

[44] Edward J. Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv:2106.09685 [cs]. Oct. 2021. DOI: 10.48550/arXiv.2106.09685. URL: http://arxiv.org/abs/2106.09685 (visited on 05/10/2025).

[45] Ilya Loshchilov and Frank Hutter. "Decoupled Weight Decay Regularization". In: *International Conference on Learning Representations (ICLR)* (2019). arXiv: 1711.05101 [cs.LG].

[46] Cheng Lu et al. *DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps*. Oct. 13, 2022. DOI: 10.48550/arXiv.2206.00927. arXiv: 2206.00927 [cs]. URL: http://arxiv.org/abs/2206.00927 (visited on 05/16/2025). Pre-published.

[47] Yaniv Taigman et al. "DeepFace: Closing the Gap to Human-Level Performance in Face Verification". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH, USA: IEEE, June 2014, pp. 1701–1708. ISBN: 978-1-4799-5118-5. DOI: 10.1109/CVPR.2014.220. URL: https://ieeexplore.ieee.org/document/6909616 (visited on 05/16/2025).

[48] Camillo Lugaresi et al. *MediaPipe: A Framework for Building Perception Pipelines*. June 14, 2019. DOI: 10.48550/arXiv.1906.08172. arXiv: 1906.08172 [cs]. URL: http://arxiv.org/abs/1906.08172 (visited on 05/16/2025). Pre-published.