DEPARTMENT OF COMPUTING SCIENCES

UNIVERSITA BOCCONI

Bachelor's Thesis in Artificial Intelligence

AnonNET: Multi-Stage Anonymized Video Synthesis Using Motion Transfer via Generative AI

Author:	Anil Egin
Supervisor:	Prof. Andrea Tangherloni
Advisor:	Dr. Antitza Dantcheva
Submission date:	20 September 2025

To my family, for their unconditional love and support.

I confirm that this bachelor's thesis is my own work and I have documented all sources and material used.

Milan, 20 September 2025

_____

Abstract

This thesis studies the task of video face anonymization, which seeks to generate new facial identities in talking-head videos while preserving natural expressions and motion. Unlike conventional face-swapping methods that replace one real identity with another, we introduce **AnonNET**, a multi-stage anonymization framework that synthesizes a completely new identity while maintaining the subject's age, gender, race, and expressions.

AnonNET first detects scene changes and extracts representative keyframes for anonymization. Using diffusion-based inpainting guided by attribute recognition and structural priors, it replaces the subject's face with a realistic synthetic alternative. To ensure temporal consistency, a landmark-free motion transfer module applies expression- and motion-aware trajectories without requiring explicit keypoint tracking. This allows the anonymized face to seamlessly mimic natural head movements, avoiding identity leakage while preserving realism.

We evaluate AnonNET on diverse datasets including VoxCeleb2, CelebV-HQ, and HDTF, which feature challenging facial dynamics. Experiments demonstrate strong identity obfuscation, high perceptual quality, and stable temporal consistency, outperforming existing face-swapping and motion transfer methods. The implementation and anonymized versions of these datasets will be publicly released to support reproducibility and further research. These results highlight AnonNET's potential for privacy-sensitive applications in journalism, law enforcement, healthcare, and beyond.

The implementation and anonymized datasets will be publicly released at `https://github.com/anilegin/AnonNET`.

Contents

1 Introduction

Contents

1.1 Motivation

Video anonymization is aimed at effectively obscuring identity-information, such as faces or voices, without compromising the integrity or usability of the content. Such anonymization has been fueled by ethical, legal or practically necessity - increasingly essential in a world, where facial images and videos have become omnipresent [9]. For instance, medical therapy sessions recorded for research require video anonymization to protect patient identities, particularly facial features, while preserving related expressions and emotions, which are pertinent for research.

In addition, legal frameworks such as the General Data Protection Regulation (*GDPR*)[1] impose strict constraints on collection, processing, and dissemination of personal data, including biometric identifiers such as images and videos of the human face. More recently, the European Union's Artificial Intelligence Act (*AI Act*))[2], adopted in 2024, introduced a tiered risk-based framework that places heightened scrutiny on AI systems handling biometric and identity-sensitive data. These evolving regulations reinforce the demand for anonymization methods that ensure privacy protection, while preserving the utility of data for downstream computer vision tasks. Traditional approaches including pixelation, blurring, and masking often degrade video quality and compromise associated applicability in such tasks [31].

[1] `https://gdpr-info.eu`

[2] `https://artificialintelligenceact.eu`

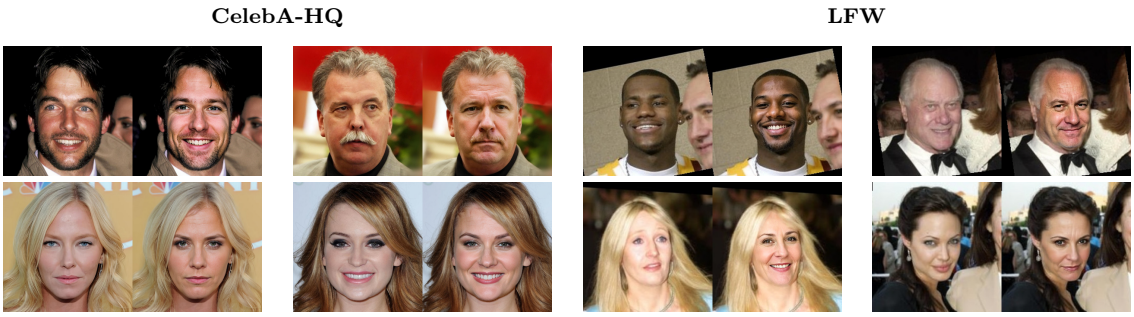

Figure 1.1: Qualitative comparison of original and anonymized faces. Columns 1–2 present examples from CelebA-HQ, and columns 3–4 present examples from LFW.

Modern deep generative models, especially Generative Adversarial Networks (GANs), have significantly advanced visual realism. *Image anonymization* inpainting-based approaches, such as DeepPrivacy [17] and DeepPrivacy2 [16] replace only sensitive facial regions, thereby preserving the surrounding content of the facial area. Conversely, fully synthetic pipelines such as FALCO [2] generate artificial facial images, while preserving high-level attributes including age, gender and race, however may introduce inconsistencies in expression or struggle with robustness under diverse pose and lighting conditions due to reliance on GAN-inversion and matching in a synthetic latent space.

W.r.t. *video anonymization*, face-swapping methods [49] can inadvertently preserve identity-specific features, compromising unlinkability [35]. Finally, landmark-based motion transfer techniques risk motion artifacts, in cases when tracking is imperfect [32].

Motivated by the above, in this work we propose *AnonNET*, a multi-stage framework that (a) *synthesizes new identities*, while preserving facial attributes. Our approach employs diffusion-based inpainting guided by structural priors for comprehensive identity obfuscation, avoiding the limitations of reference-based methods. Further, (b) a landmark-free motion transfer module ensures realistic expressions without relying on explicit keypoint tracking, thereby mitigating alignment fail cases. By restricting modifications to the face region, AnonNET retains original scene context, offering high-quality, privacy-preserving video anonymization.

1.2 Contributions

The main contributions of this thesis include the following.

- A *novel multi-stage framework* for video anonymization that synthesizes new facial identities while preserving age, gender, race, and expressions.
- A *new dataset of anonymized videos* pertained to VoxCeleb, CelebV, and HDTF datasets, providing a valuable resource for future research in areas such as deepfake detection.
- A *comprehensive evaluation* of our pipeline against state-of-the-art models on *image level*, as well as with regard to the re-identification, identity consistency, and expression-aware downstream utility on *video level*.

2
Related Work

Contents

2.1 Face Anonymization

Traditional Techniques. Classical anonymization approaches such as pixelation, blurring, and masking have been widely used to obscure facial features while maintaining general image context. Their simplicity makes them appealing, yet they significantly distort images, impeding downstream tasks such as facial recognition, expression analysis, and attribute prediction [3, 27, 19]. Moreover, several works have shown that these techniques can be partially reversible [24, 7], raising concerns about robustness and privacy leakage. Consequently, traditional methods are insufficient for high-security or analytic scenarios that demand both privacy protection and visual utility.

Adversarial Techniques. A line of work employs *adversarial training*, balancing anonymization and utility preservation in a min-max optimization framework. Nasr et al. [26] introduced adversarial regularization to mitigate membership inference, while Wu et al. [42, 41] leveraged GANs to de-identify faces without compromising action recognition. However, training GANs directly in image space is challenging, as fine-grained details such as pose, emotion, and background are difficult to preserve under high pixel-space complexity. To address this, some methods operate in latent spaces, disentangling identity from other factors. For instance, StyleID [22] manipulates latent codes to alter identity traits, yet may inadvertently preserve bias-inducing features. GAN-based inpainting approaches [34, 14, 23] rely on landmarks or segmentation masks to guide anonymization, but failures in auxiliary data can compromise both anonymity

and realism. DeepPrivacy2 [16] extended these frameworks to full-body anonymization, but remained dependent on precise segmentation, making it vulnerable to landmark errors.

Diffusion-Based Approaches. Diffusion models refine noisy inputs through iterative denoising, often with U-Net architectures, and have recently been applied to anonymization [20, 28, 21]. While these methods produce visually convincing results, they typically apply fixed or narrowly parameterized transformations, limiting diversity in synthesized identities. This makes it difficult to consistently control attributes such as age, gender, race, and expressions. In contrast, our approach leverages diffusion-based inpainting conditioned on explicit attribute priors, providing fine-grained control over non-identifying characteristics while robustly obscuring identity. This ensures a wider variety of anonymized appearances, retains scene context, and enables downstream tasks such as emotion recognition or pose estimation without significant degradation.

2.2 Motion Transfer and Video Re-Animation

Motion transfer seeks to animate faces or bodies from a reference image, guided by motion cues such as head pose or expressions [44, 13]. Early works relied on explicit structural priors—keypoints, semantic parsing, or 3D models [38]—which are prone to errors under occlusions or complex poses. More recent methods operate in latent spaces, relaxing dependence on landmarks and offering greater flexibility [11].

Landmark-Based Methods. Techniques such as FOMM, MRAA, SAFA, and Face vid2vid [33, 48, 36, 37] estimate 2D/3D keypoints or deformation fields and warp source images accordingly. While capable of high-quality results with accurate landmarks, these pipelines degrade under large motions or misdetections. Some integrate face-swapping, risking unintended modifications to background or non-identity features. Landmark-based methods thus struggle with extreme head rotations, fine-grained expressions, and multi-person scenarios.

Landmark-Free Methods. Latent-space approaches synthesize motion directly, capturing subtle expressions without explicit structural priors. LIA (*Latent Image Animator*) [38] learns a motion dictionary in latent space, while LivePortrait [11] introduces stitching and retargeting controls to improve robustness to varied poses. These methods achieve smoother motion transfer and reduce reliance on landmark accuracy, though challenges remain for extreme occlusions or high-resolution details. In our work, we combine LIA's latent dictionary with LivePortrait's retargeting to achieve consistent motion transfer with improved visual fidelity.

Additional Models. Recent frameworks such as FADM [48], Face Adapter [12], AniPortrait [39], X-Portrait [43], and MegActor [46] achieve high-quality reenactment but often rely on 3D priors, explicit landmarks, or heavy computational resources. These trade-offs limit their applicability in large-scale or real-time video anonymization.

2.3 Comparison of Existing Pipelines

End-to-end pipelines have sought to unify anonymization and motion transfer. **RID-TWIN** [25] used BLIP for face captioning and stable diffusion jointly with MediaPipe-based segment extraction. However, it focused on automatic de-identification rather than preserving specific attributes like age, gender, race, or expressions. SAFA was leveraged [36] for head motion transfer, which is prone to landmark-based errors. SAFA used self-supervised landmark-like keypoints, however associated low resolution leads to motion and appearance artifacts.

AI Stylization [45] constitutes a perceptual approach for anonymization, replacing facial realism with *artistic* abstraction. After an initial facial feature randomization stage, the method applied cubist and painterly stylizations to anonymized faces, aiming to preserve emotional salience and enhance viewer empathy. However, the approach sacrificed photorealism entirely, as renderings tend to be stylized and visually inconsistent across frames.

In contrast, our **AnonNET** framework conditions diffusion-based inpainting on user-defined attribute priors, enabling consistent preservation of essential characteristics. Specifically, we adopt the *landmark-free* motion transfer framework LIA and LivePortrait, in order to transfer expression and head pose. Additionally, we systematically detect *scene changes* for improved temporal consistency and seamless anonymization across longer video sequences, therefore accommodating complex multi-scene videos effectively.

3 Method

Contents

3.1 Overview of the Proposed Framework

We propose a *multi-stage* pipeline (see Figure 3.1), streamlined to obfuscate identity, while preserving attributes such as age, gender, race, and at the same time ensuring temporal consistency represented by expression and head poses. In particular, our AnonNET includes following stages.

(1) Scene Detection & Identity Clustering. We segment the input video via FFmpeg-based scene change detection, and then cluster faces across scenes using VGG-Face2 [5] embeddings and cosine-distance thresholds. Scenes containing the same individual share a consistent anonymized identity throughout the video.

(2) Face Detection & Frontal Selection. RetinaFace [10] localizes the face region. A single representative frame, associated to a frontal pose, is selected per scene. This reduces flickering and computational burden by focusing anonymization on a single frame per segment.

(3) Attribute Recognition. We estimate age, gender, race, and emotion via DeepFace [30], retaining high-level features that do not reveal identity.

(4) Diffusion-Based Inpainting. We adopt Realistic Vision V5.0 to inpaint the masked face guided by **ControlNets** for segmentation mask, lineart, and openpose maintain structural fidelity along with **Attribute-Conditioned Prompt** that we provide key attributes while discarding identity-specific details.

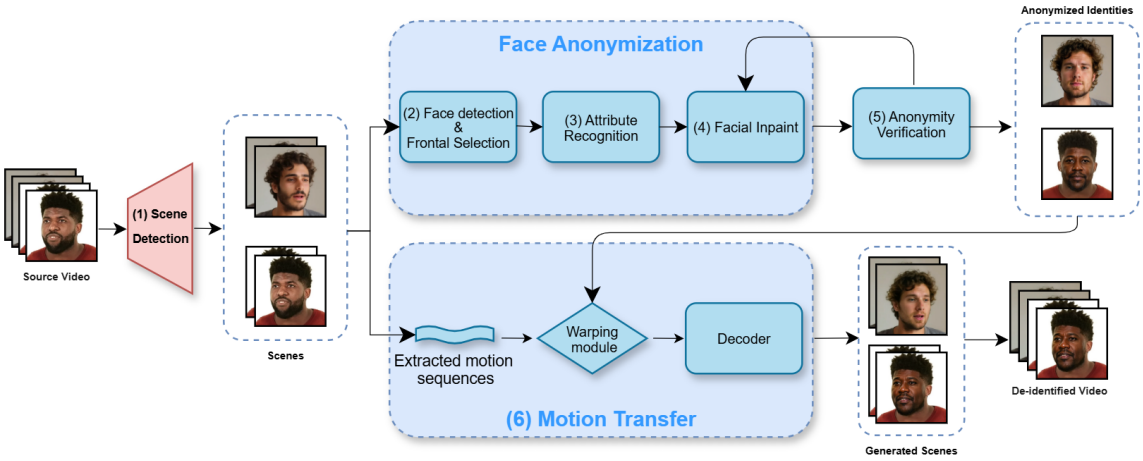

Figure 3.1: Overview of our multi-stage anonymization AnonNET-pipeline. (1) Scene changes are detected, and identities are tracked. (2) Faces are detected and single frontal frame is selected per scene identity. (3) Facial attributes are recognized. (4) A diffusion-based model inpaints the masked face. (5) Current anonymity is evaluated. (6) Landmark-free motion transfer reintroduces natural head movement. (7) Frames are reassembled for a coherent output video.

(5) Anonymity Verification. To ensure identity obfuscation and avoid leakage at the end of the process, we include a verification module that verifies the cosine similarity between VGG-Face2 [5] embeddings of the original and anonymized images. In case that the similarity exceeds a threshold, the inpainting is re-triggered with higher stochasticity to enforce stronger anonymization.

(6) Landmark-Free Motion Transfer. We select the frameworks **LIA** and **LivePortrait** to warp the anonymized face per frame, replicating natural head movements from the original video, without explicit landmark tracking. This approach mitigates flickering and alignment errors.

(7) Video Reassembly. Processed frames are merged back into the original scene structure, retaining audio and background context.

We note that resulting videos *preserve facial attributes, exhibit strong identity obfuscation, and temporal coherence*, see Supplementary Material for videos. Our modular design allows for each stage, namely scene detection, face detection, inpainting, motion transfer to be independently improved or replaced. We proceed to elaborate on each stage.

3.2 Video Preprocessing and Scene Detection

(1) Scene Change Detection. We detect coarse scene boundaries using FFmpeg's scene change filter (`select='gt(scene,X)'`), which flags transitions based on frame-wise histogram differences. To avoid over-segmentation, we refine these segments by computing mean RGB differences and merging visually similar intervals under a shared `scene_id`. This simple yet effective two-stage strategy yields stable scene partitions and reduces redundancy, enabling consistent identity tracking and efficient anonymization across temporally coherent regions.

The algorithm ensures consistent identity assignment across scenes by leverag-

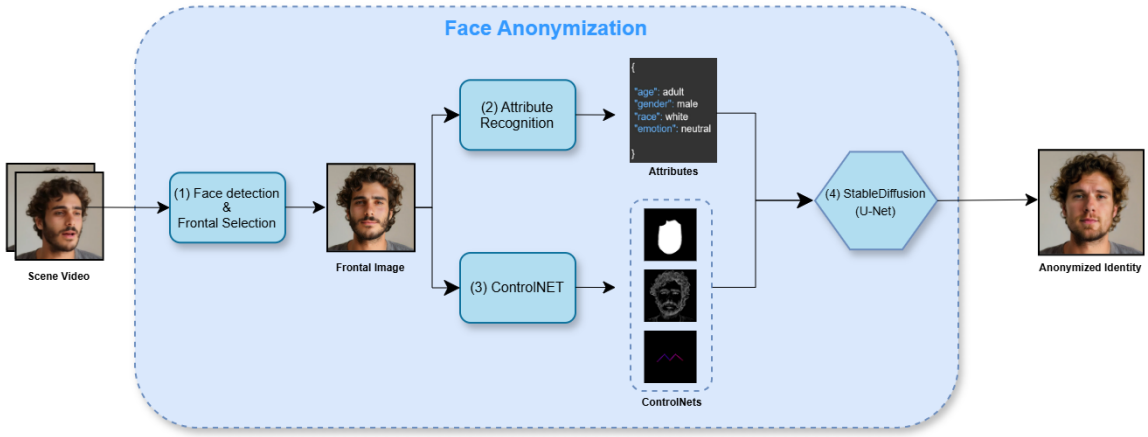

Figure 3.2: Overview of the expression-consistent face anonymization module in AnonNET. (1) A face detection and frontal selection stage extracts a frame containing a frontal face from the input scene video. Then this frontal image is processed in parallel by two branches: (2) An attribute recognition module that infers semantic attributes such as age, gender, race, and expression; (3) a ControlNet module, which extracts structural guidance (*e.g.,* face mask, lineart, pose) for conditioning the generative model. (4) Stable Diffusion based on U-Net synthesizes an anonymized face conditioned on both, extracted attributes and ControlNet features.

ing both visual similarity and embedding-based clustering. This allows the same anonymized identity to be preserved for recurring individuals throughout the video, which is critical for maintaining coherence in temporally or contextually linked segments.

(2) Frontal Selection For each scene, we select a *single* representative frame to serve as the anchor for anonymization and motion transfer. To ensure full facial coverage and minimize downstream hallucination, we prioritize frames with a frontal head pose, where both geometric structure and semantic attributes are fully visible.

We estimate head pose using the `face_alignment` [4] library in 2D landmark mode, extracting 68 facial landmarks per frame. A subset of six key points (nose tip, chin, eye corners, mouth corners) is selected and matched to a predefined 3D face model. We then

Algorithm 1 Scene Detection and Identity Clustering

Input: Video V, thresholds $\theta_{ffmpeg}, \theta_{sim}, \theta_{embed}$
Output: Scene list $\mathcal{S}$
Detect scene cuts using FFmpeg (θ_{ffmpeg})
for each scene (s, e) **do**
    Reuse or assign scene_id based on RGB diff ($< \theta_{sim}$)
    for each face embedding **do**
        Assign identity if distance $< \theta_{embed}$; else create new
    end for
    Append $(s, e, scene\_id, identities)$ to $\mathcal{S}$
end for
return $\mathcal{S}$

solve the perspective-n-point (PnP) problem via OpenCV's `solvePnP`, computing the 3D rotation vector of the head. Frames with fewer than 80% of landmarks falling within image boundaries are discarded. Among valid candidates, the frame with the smallest absolute pitch and yaw is selected as the scene frontal frame. This selection strategy ensures that identity obfuscation operates on a complete and unobstructed face. Since motion transfer is applied post-anonymization, any missing or occluded facial regions in the frontal frame would otherwise be synthesized without constraint—potentially leading to artifacts or semantic drift.

3.3 Face Detection and Attribute-Guided Prompt Generation

(3) Attribute Recognition. The localized face is then passed to the DeepFace library [30], in order to extract coarse demographic and affective attributes, including age, gender, race, and emotion. These attributes are used to guide the anonymization process in a non-identifying manner.

3.4 Diffusion-Based Inpainting

(4) Diffusion-Based Inpainting. Towards conditioning the diffusion-based inpainting, we construct a descriptive prompt, encoding the extracted attributes. Additionally, we apply a negative prompt to suppress undesired artifacts such as distortions, unrealistic textures, or cartoon-like features. The final prompt directs the generation toward a photorealistic, high-fidelity identity with preserved semantic traits. An example prompt is:

```
A photorealistic portrait of a middle-aged Asian female, with a
neutral expression.
```

We perform identity obfuscation via latent diffusion-based inpainting using Realistic Vision V5.0, a publicly available checkpoint based on Stable Diffusion v1.5 [29].

ControlNet applies structural guidance from the head mask to confine synthesis to the face region while preserving surrounding content.

Lineart & OpenPose ControlNets provide edge and pose priors to enforce geometric and expression fidelity during generation.

We use the DPMSolver++ scheduler for efficient denoising, typically over `20-70` steps with a guidance scale between `8-20`, depending on dataset characteristics. A VAE with perceptual reconstruction loss is used to map images between pixel and latent space, supporting visually coherent and detailed inpainting.

3.5 Anonymity Verification and Motion Transfer

(5) Anonymity Verification. To ensure successful identity obfuscation, we capture the cosine similarity between the VGG-Face2 embeddings of original and anonymized images. In case that the cosine distance score is below a threshold of 0.3, indicating potential identity leakage, we re-trigger the inpainting process. In this second pass, we introduce greater stochasticity by increasing the prompt guidance scale and reducing

ControlNet conditioning strength. Additionally, we extend the number of denoising steps by 5, in order to allow the diffusion process to deviate further from the original identity while still maintaining attribute consistency.

(6) Motion Transfer. We combine landmark-free frameworks (**LIA**, **LivePortrait**) with scene stitching and eye/lip retargeting to animate anonymized faces across frames:

1. **Encoding:** The original (non-anonymized) source frame is encoded to obtain a latent motion representation $z_{s \to r}$, capturing pose and expression dynamics.
2. **Framewise Motion Codes:** For each target frame, motion offsets $\mathbf{w}_{r \to d}$ are predicted based on pose and expression changes.
3. **Flow Field Synthesis:** The anonymized source frame image is encoded, and the learned flow map (from $z_{s \to r} + \mathbf{w}_{r \to d}$) is used to warp it, transferring motion to the anonymized face, while preserving appearance.
4. **Refinement:** LivePortrait optionally enhances eye and lip dynamics (*e.g.,* blinks, speech) for improved realism.

By decoupling motion from identity and operating in latent and flow-guided spaces, these approaches avoid explicit landmark detection, reduce artifacts, and increase robustness to rapid motion and partial occlusion, while maintaining temporal consistency.

(5) Video Reassembly. Finally, we reassemble processed scenes into the final video. Specifically, each anonymized segment is integrated using original timestamps, ensuring proper alignment. We note that the audio is unaltered and resynchronized to preserve speech and background sounds.

Scenes with no faces remain untouched; multi-person scenarios are deferred for future work due to the complexities of simultaneous multi-face motion transfer.

4 Results

Contents

4.1 Experimental Setup

We proceed to comprehensively evaluate AnonNET's performance and compare related results to state-of-the-art anonymization methods, providing quantitative analysis on identity obfuscation and attribute retention. Further, an ablation study illustrates the impact of core components.

Datasets

We evaluate our framework on following two widely used *image* datasets.

CelebA-HQ [18] contains 30,000 high-resolution face images annotated with 40 facial attributes, including age, gender, race, and appearance traits. The diversity in pose and lighting allow for our evaluation on attribute-preserving anonymization.

LFW [15] includes 13,233 images of 5,749 identities captured in unconstrained conditions. We focus on identity obfuscation and generalization under varying image quality and occlusions.

For *video*-based anonymization, we additionally use following datasets.

CelebV-HQ [51] constitutes a curated high-resolution video face dataset.

VoxCeleb2 [8] represents a subset of 50,000 clips featuring over thousands of identities in varied conditions.

HDTF [50] comprises expressive head motion and fine-grained lip sync.

Comparative Methods

We compare AnonNET against following *image anonymization* frameworks. DeepPrivacy2 [16] is prominent for removing identity cues, while preserving contextual and structural consistency. At the same time CIAGAN [23] represents a competitive former approach that modifies latent identity features while retaining key facial attributes.

In addition, *w.r.t. video-anonymization*, we compare AnonNET to RID-TWIN [25] that constitutes an end-to-end video anonymization approach with temporal coherence.

Implementation Details

AnonNET integrates a multi-stage pipeline with tailored configurations per component:

Anonymization. Based on Realistic Vision V5.0 with:

- Denoising steps: 20–70 (dataset-specific tuning)
- Guidance scale: 8–20 (for prompt expressiveness)
- ControlNets: Segmentation, LineArt, and OpenPose
- DPMSolver Scheduler for accelerated sampling

Motion Transfer. Landmark-free video animation via LivePortrait [11] and LIA [38], enabling smooth expression preservation across frames.

Computational Time. Anonymizing the 50,000-clip VoxCeleb2 subset takes approximately 160 hours with LIA and 185 hours with LivePortrait, using a single A100 GPU.

4.2 Evaluation Metrics

We adopt an evaluation framework, aimed at assessing identity obfuscation, attribute retention, and visual quality.

Re-identification Rate (Re@1). To evaluate identity leakage, we measure the rank-1 re-identification accuracy using face embeddings extracted from **VGGFace2** [5] and **CASIA-WebFace** [47] models. For each anonymized image, we compute its embedding and retrieve the closest match from the original image set based on cosine similarity. A sample is considered successfully re-identified if its nearest neighbor corresponds to the same identity as the original. The final Re@1 score is computed as the ratio of correctly re-identified samples over the total number of anonymized images. Lower scores indicate stronger identity obfuscation.

Image Quality and Aesthetics. We use the **Q-Align/One-Align** [40] metric to estimate perceptual quality (Qual) and visual appeal (Aes). These scores are averaged across all images and videos.

Pose and Gaze Preservation. For each image, facial landmarks are first localized with MTCNN. Pose angles (pitch/yaw) are extracted via Dlib's face pose estimator. Gaze direction is then evaluated using **L2CS-Net** [1], computing the mean absolute error (MAE) between original and anonymized outputs.

Expression Preservation. Expression labels are predicted pre- and post-anonymization using the **DeepFace** library [30]. Accuracy is defined as the fraction of samples, where the predominant expression label is retained.

Figure 4.1: Qualitative face anonymization results pertained to the CelebA-HQ dataset. Each row corresponds to an input image (left column), and columns show outputs from various image-anonymization methods.

Temporal Identity Consistency (Video). For each anonymized video, we compute the average cosine distance of DINO [6] embeddings between consecutive frames, comparing with the same metric on original videos. This assesses intra-video consistency post-anonymization.

4.3 Quantitative Evaluation

Re-identification Performance. Table 4.1 reports rank-1 re-identification accuracy (Re@1) employing VGGFace2 and CASIA-WebFace embeddings on CelebA-HQ and LFW. While CIAGAN and DeepPrivacy2 achieve the lowest scores overall, AnonNET (35 steps) remains competitive, with Re@1 values of 0.041 (VGG) on CelebA-HQ and 0.042 on LFW. Compared to CIAGAN and DeepPrivacy, AnonNET provides a consistent drop in re-identification, while retaining attribute fidelity and performs favorably relative to recent diffusion-based anonymization baselines such as FALCO and CAMOUFLaGE.

Perceptual Quality and Aesthetic Appeal. As shown in Table 4.2, AnonNET outperforms all baselines on Q-Align quality and aesthetic scores across both datasets. *W.r.t.* CelebA-HQ, it achieves the highest quality (4.164) and aesthetics (3.332) scores, exceeding both, DeepPrivacy2 and ground truth. *W.r.t.* LFW, AnonNET similarly leads with 2.887 (Qual) and 1.939 (Aes), indicating strong generalization to unconstrained, low-resolution data. In contrast, CIAGAN yields lower perceptual scores

Encoding	CelebA-HQ		LFW	
	VGG↓	CASIA↓	VGG↓	CASIA↓
DeepPrivacy2 [16]	0.008	**0.008**	0.023	0.017
DeepPrivacy [17]	0.011	0.036	0.015	0.027
CIAGAN [23]	**0.004**	0.022	**0.009**	**0.002**
FALCO [2]	0.017	0.028	0.016	0.021
CAMOUFLaGE [28]	0.096	0.100	0.102	0.116
AnonNET *(steps = 20)*	0.073	0.031	0.056	0.039
AnonNET *(steps = 35)*	0.041	0.017	0.042	0.027

Table 4.1: Re-identification rate employing VGGFace2 and CASIA pertaining to the CelebA-HQ [18] and LFW [15] datasets. Lower scores denote a lower similarity between anonymized and original images and are therefore better.

than both, AnonNET and DeepPrivacy2, consistent with degradation as noted in prior results. Beyond preserving visual realism, AnonNET systematically reduces artifacts and low-quality regions, owing to its attribute-guided diffusion design. These results are coherent and photorealistic outputs enhance compatibility with downstream tasks such as expression recognition or affective computing.

Figure 4.2: Qualitative comparison of original and anonymized frames pertained to the VoxCeleb2 dataset using AnonNET. Each column shows original/anonymized pairs.

Encoding	CelebA-HQ		LFW	
	Qual↑	Aes↑	Qual↑	Aes↑
GT	4.035	2.932	2.047	1.191
DeepPrivacy2	3.551	1.875	2.025	1.099
CIAGAN	1.011	1.361	1.006	1.466
AnonNET *(steps = 20)*	4.074	3.055	**2.914**	1.904
AnonNET *(steps = 35)*	**4.164**	**3.332**	2.887	**1.939**

Table 4.2: Quality and aesthetics scores for anonymized images of CelebA-HQ and LFW.

Trade-off. The above results confirm that AnonNET offers a privacy–utility trade-off: while not achieving the absolute lowest Re@1, it provides superior image quality.

Video-level Evaluation. Table 4.3 compares identity preservation, perceptual quality, and aesthetics between ground truth videos and those anonymized by AnonNET. Across all three datasets, AnonNET improves both, quality and aesthetics scores over

the original videos while maintaining comparable levels of identity suppression. On CelebV-HQ and HDTF, our method achieves higher quality (*e.g.,* 4.153 versus 4.045 on HDTF) and aesthetics (*e.g.*, 3.021 versus 2.938), with only marginal differences in identity preservation.

W.r.t. VoxCeleb2, which encompasses low resolution and challenging visual settings in related videos, AnonNET produces clean and coherently anonymized faces, raising quality from 2.493 to 2.859 and aesthetics from 1.606 to 1.949. These results highlight the robustness of our pipeline, even in unconstrained settings. Indeed, the new synthesized face improves structural integrity and overall visual consistency, rendering anonymized videos amenable to downstream tasks such as tracking or expression analysis.

Dataset	**GT**			**AnonNET**		
	id_pres ↓	qual ↑	aes ↑	id_pres ↓	qual ↑	aes ↑
CelebV-HQ	**0.011**	3.800	2.718	0.013	**3.907**	**2.886**
VoxCeleb2	**0.021**	2.493	1.606	0.022	**2.859**	**1.949**
HDTF	0.008	4.045	2.938	**0.007**	**4.153**	**3.021**

Table 4.3: Comparison of identity preservation, quality, and aesthetics in videos for ground truth and AnonNET across datasets. Lower is better for identity preservation (id_pres); higher is better for quality (qual) and aesthetics (aes).

(a) Gender mismatch

(b) Expression mismatch

(c) Gaze inconsistency

Figure 4.3: Comparison of original and anonymized images illustrating limitations of the proposed AnonNET framework.

As shown in Figure 4.2, AnonNET reconstructs sharper facial features and preserves expressions more consistently than the original VoxCeleb2 frames, which suffer from heavy blur and compression. This visual improvement, especially in motion-rich regions, renders previously unusable videos viable for downstream tasks such as expression analysis or video reenactment.

Table 4.4 shows that AnonNET, guided by OpenPose, achieves superior pose preservation and competitive gaze alignment. Even with fewer denoising steps, it outperforms all baselines, highlighting its efficiency and motion consistency.

Table 4.5 summarizes AnonNET's performance on CelebA-HQ and LFW. our proposed method achieves near-perfect anonymization rates with no detection failures. Attribute preservation remains high across both datasets, particularly for gender and

Dataset	CelebA-HQ	
	Pose↓	Gaze↓
DeepPrivacy2	0.140	0.244
FALCO	0.088	0.258
FAMS[21]	0.048	**0.161**
AnonNET *(steps = 20)*	**0.014**	0.187
AnonNET *(steps = 35)*	0.015	0.172

Table 4.4: Pose and gaze preservation (lower is better) on CelebA-HQ.

race, while expression accuracy is lower on LFW due to its greater variability and resolution constraints. These results confirm AnonNET's robustness across datasets with differing visual and demographic characteristics.

Metric	CelebA-HQ	LFW
Total images	30,000	13,233
Successfully anonymized	29,997	12,912
Anonymization failures	3	321
Face detection failures	0	0
Race (%)	79.5	87.1
Gender (%)	99.4	99.3
Age (mean ± std)	(1.87, 4.23)	(2.69, 6.29)
Expression (%)	74.7	52.9

Table 4.5: Anonymization statistics and attribute preservation accuracy on CelebA-HQ and LFW datasets.

Limitations. Figure 4.3 highlights failure cases of AnonNET. Since attribute guidance relies on pretrained recognition networks, errors in gender, expression, or gaze estimation can propagate to the anonymized output. These limitations suggest the need for more robust or fine-tuned attribute predictors, especially for handling edge cases and underrepresented demographics.

5 Conclusions

In this work, we introduced AnonNET, a unified multi-stage framework for anonymizing talking head videos, placing emphasis on preserving key facial attributes. We presented extensive evaluations, demonstrating the ability of AnonNET to obfuscate identity, while allowing for further analysis and downstream task utility. It offers competitive re-identification resistance, while significantly outperforming prior methods in visual quality, aesthetics, and attribute preservation.

As opposed to the state of the art, AnonNET is robust to diverse poses, lighting conditions, and motion dynamics, rendering it suitable for real-world applications such as journalism, therapy, and human-computer interaction. We place emphasis on high visual quality and temporal consistency. Specifically, our framework combines diffusion-based inpainting and landmark-free motion transfer that jointly create photorealistic anonymized faces that are expression-consistent and structurally aligned across video frames.

Through extensive evaluations on several benchmark datasets, we demonstrated that AnonNET achieves a favorable trade-off between identity obfuscation and downstream task utility. It offers competitive re-identification resistance, while significantly outperforming prior methods in terms of visual quality, aesthetics, and attribute preservation. Compared to earlier techniques, AnonNET is more robust to diverse poses, lighting conditions, and motion dynamics, rendering it suitable for real-world applications such as journalism, therapy, and human-computer interaction.

6 Future Work

Contents

6.1 Real-Time or Near Real-Time Processing

One promising direction for future research involves extending our anonymization pipeline to operate in real-time or near real-time environments. Currently, the system's computational complexity and reliance on multiple pretrained models result in significant processing delays, particularly with high-resolution videos or extensive frame sequences. Future efforts could explore optimization techniques such as model distillation, quantization, or efficient architectures specifically tailored for edge devices. Additionally, leveraging GPU-accelerated inference or parallelization strategies may further decrease latency, enabling practical applications in live broadcasting, video conferencing, and streaming contexts.

6.2 Group Interactions and Multi-Person Scenes

Presently, our motion transfer methods primarily support single-person facial animations due to limitations inherent to the underlying pretrained animation models. Extending anonymization to reliably handle multi-person interactions and complex group scenes remains a critical challenge. Future developments should investigate advanced multi-agent tracking methods, sophisticated head segmentation algorithms, and robust identity-consistent animation frameworks capable of managing simultaneous motions and interactions across multiple identities. Enhancements in scene understanding, such as modeling social interactions, spatial dynamics, and attention mechanisms, could also significantly improve realism and consistency in anonymized group scenarios.

6.3 Improving Diversity and Control of Generated Identities

A key area for future research involves enhancing the diversity and fine-grained controllability of generated anonymous identities. Although our approach currently provides basic control over age, gender, race, and expression attributes, further research could explore richer latent space manipulation techniques and conditioning mechanisms. By developing methods for explicit user-defined constraints, personalized anonymization styles, or adaptive prompts based on semantic scene context, the anonymization process could achieve greater variability and coherence. Furthermore, incorporating fairness-aware generative modeling and systematic bias mitigation approaches would ensure equitable representation across diverse demographic groups, enhancing the ethical robustness of the anonymization outcomes.

Bibliography

[1] Ahmed A. Abdelrahman, Thorsten Hempel, Aly Khalifa, and Ayoub Al-Hamadi. L2cs-net: Fine-grained gaze estimation in unconstrained environments. *arXiv preprint arXiv:2203.03339*, 2022.

[2] Simone Barattin, Christos Tzelepis, Ioannis Patras, and Nicu Sebe. Attribute-preserving face dataset anonymization via latent code optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8001–8010, 2023.

[3] M. Barrett et al. Face anonymization: A comparative study of traditional methods. *Journal of Privacy and Security*, 4:123–145, 2017.

[4] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017.

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[7] S. Choi et al. Blurring and masking for face obfuscation: A comprehensive review. *Security and Privacy Research*, 8:233–250, 2021.

[8] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

[9] Antitza Dantcheva. *Computer vision for deciphering and generating faces*. Habilitation thesis, Université Côte d'Azur, September 2021.

[10] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020.

[11] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024.

[12] Yue Han, Junwei Zhu, Keke He, Xu Chen, Yanhao Ge, Wei Li, Xiangtai Li, Jiangning Zhang, Chengjie Wang, and Yong Liu. Face-adapter for pre-trained diffusion models with fine-grained id and attribute control. In *European Conference on Computer Vision*, pages 20–36. Springer, 2024.

[13] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024.

[14] Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, and Libing Wu. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15014–15023, 2022.

[15] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[16] Håkon Hukkelås and Frank Lindseth. Deepprivacy2: Towards realistic full-body anonymization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1329–1338, 2023.

[17] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In *International Symposium on Visual Computing (ISVC)*, pages 565–578. Springer, 2019.

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[19] L. Kramer et al. Pixelation as a privacy-preserving technique in facial recognition systems. *International Journal of Privacy Computing*, 12:34–46, 2019.

[20] Han-Wei Kung, Tuomas Varanka, Sanjay Saha, Terence Sim, and Nicu Sebe. Face anonymization made simple. *arXiv preprint arXiv:2411.00762*, 2024.

[21] Han-Wei Kung, Tuomas Varanka, Sanjay Saha, Terence Sim, and Nicu Sebe. Face anonymization made simple. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 1040–1050, February 2025.

[22] Minh-Ha Le and Niklas Carlsson. Styleid: Identity disentanglement for anonymizing faces. *arXiv preprint arXiv:2212.13791*, 2022.

[23] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5447–5456, 2020.

[24] Richard McPherson, Reza Shokri, and Vitaly Shmatikov. Defeating image obfuscation with deep learning. *arXiv preprint arXiv:1609.00408*, 2016.

[25] Anirban Mukherjee, Monjoy Narayan Choudhury, and Dinesh Babu Jayagopi. Rid-twin: An end-to-end pipeline for automatic face de-identification in videos. *arXiv preprint arXiv:2403.10058*, 2024.

[26] Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, pages 634–646, 2018.

[27] Elaine M Newton, Latanya Sweeney, and Bradley Malin. Preserving privacy by de-identifying face images. *IEEE transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005.

[28] Luca Piano, Pietro Basci, Fabrizio Lamberti, and Lia Morra. Latent diffusion models for attribute-preserving image anonymization. *arXiv preprint arXiv:2403.14790*, 2024.

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

[30] Sefik Serengil and Alper Ozpinar. A benchmark of facial recognition pipelines and co-usability performances of modules. *Journal of Information Technologies*, 17(2):95–107, 2024.

[31] Yan Shoshitaishvili, Christopher Kruegel, and Giovanni Vigna. Portrait of a privacy invasion. *Proceedings on Privacy Enhancing Technologies*, 2015(1):41–60, 2015.

[32] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Neural Information Processing Systems (NeurIPS)*, 2019.

[33] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13653–13662, 2021.

[34] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and effective obfuscation by head inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5050–5059, 2018.

[35] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Neural texture synthesis and expression transfer for face video reenactment. In *IEEE Transactions on Visualization and Computer Graphics*, 2019.

[36] Qiulin Wang, Lu Zhang, and Bo Li. Safa: Structure aware face animation. In *2021 International Conference on 3D Vision (3DV)*, pages 679–688. IEEE, 2021.

[37] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021.

[38] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*, 2022.

[39] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024.

[40] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

[41] Yifan Wu, Fan Yang, Yong Xu, and Haibin Ling. Privacy-protective-gan for privacy preserving face de-identification. *Journal of Computer Science and Technology*, 34:47–60, 2019.

[42] Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In *Proceedings of the European conference on computer vision (ECCV)*, pages 606–624, 2018.

[43] You Xie, Hongyi Xu, Guoxian Song, Chao Wang, Yichun Shi, and Linjie Luo. X-portrait: Expressive portrait animation with hierarchical motion attention. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[44] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024.

[45] Özge Nilay Yalçın, Vanessa Utz, and Steve DiPaola. Empathy through aesthetics: Using ai stylization for visual anonymization of interview videos. In *Proceedings of the 3rd Empathy-Centric Design Workshop: Scrutinizing Empathy Beyond the Individual*, pages 63–68, 2024.

[46] Shurong Yang, Huadong Li, Juhao Wu, Minhao Jing, Linze Li, Renhe Ji, Jiajun Liang, and Haoqiang Fan. Megactor: Harness the power of raw video for vivid portrait animation. *arXiv preprint arXiv:2405.20851*, 2024.

[47] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[48] Bohan Zeng, Xuhui Liu, Sicheng Gao, Boyu Liu, Hong Li, Jianzhuang Liu, and Baochang Zhang. Face animation with an attribute-guided diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 628–637, 2023.

[49] W. Zhang, S. Shan, and X. Chen. Faceshifter: Towards high fidelity and occlusion aware face swapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[50] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 3661–3670, 2021.

[51] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pages 650–667. Springer, 2022.

A
Appendix

Contents

A.1 Implementation Details

Environment and Dependencies

Hardware. All experiments were conducted on an NVIDIA A100 (40 GB). We verified that the full pipeline also runs on an NVIDIA Tesla V100 (16 GB). For video mode with default settings we recommend at least **15 GB** of GPU memory.

Core libraries. The implementation relies on `PyTorch`, `diffusers`, `controlnet_aux`, `onnxruntime` (for YOLO in ONNX), `opencv-python`, `DeepFace`, and `FFmpeg`.

Checkpoints and models.

- *Stable Diffusion:* `SG161222/Realistic_Vision_V5.0_noVAE`;
  VAE: `sd-vae-ft-mse-original` (FP16).
- *ControlNets:* inpaint, lineart, and openpose heads from the SD 1.5 family.
- *Segmentation:* custom head-segmentation model, derived from BiSeNet (ResNet–18) face parsing.
- *Detectors:* custom YOLOv8 (ONNX) and RetinaFace (via DeepFace).
- *Motion transfer:* LivePortrait (default) and LIA (Latent Image Animator; weights `vox.pt`).

Schedulers/precision. DPMSolverMultistepScheduler; FP16 throughout; xFormers enabled when available.

Quick Start

Image → anonymized image

```
python image_anonymize.py --image input.jpg --segment face \
  --steps 35 --strength 0.9 0.4 0.3 --guidance_scale 8.0 \
  --max_height 612 --max_width 612 --negative_prompt "" \
  --save_folder ./results
```

Writes `./results/input_anonymized_*.png`.

Video → anonymized (LivePortrait, default)

```
python anonymize.py --driving_path input.mp4 --motion lp \
  --save_folder ./results --scene_threshold 0.2 \
  --scene_similarity_threshold 2.0 --segment face \
  --max_height 612 --max_width 612 --negative_prompt ""
```

Produces `_output_lp.mp4` and `_output_lp_concat.mp4`.

Large head motion (disable stitching)

```
python anonymize.py --driving_path long.mp4 --motion lp \
  --no_stitch --save_folder ./results --segment head
```

Alternative motion backend (LIA)

```
python anonymize.py --driving_path input.mp4 --motion lia \
  --save_folder ./results
```

Top-level Pipeline and CLI

Video anonymization

```
python anonymize.py --driving_path <video.mp4> \
  --save_folder ./results --motion {lp|lia} \
  --scene_threshold 0.2 --scene_similarity_threshold 2.0 \
  --segment {face|head} --max_height 612 --max_width 612 \
  --negative_prompt "" [--no_stitch] [--max_len <sec>] [--cache]
```

Image anonymization

```
python image_anonymize.py --image <img.png> --segment {face|head} \
  --steps 35 --strength 0.9 0.4 0.3 --guidance_scale 8.0 \
  --max_height 612 --max_width 612 --negative_prompt "" \
  [--seed <int>] --save_folder ./results
```

Video Preprocessing and Scene Handling

Scenes are detected using `detect_scenes_and_assign_ids` with `scene_threshold`=0.2 and `scene_similarity_threshold`=2.0. Keyframes are selected via `get_frontal_frame` (`best`=True). Scene-level identities are linked using `DeepFace.verify` (VGG-Face, cosine). New identities are assigned a random seed in $[1000, 9999]$; associated attributes (age, gender, race, emotion) are cached for consistency across scenes.

Detection, Segmentation, and Masking

The primary detector is YOLOv8 (ONNX) with `conf`=0.8 and `IoU`=0.3. Bounding boxes are square-padded with a margin; RetinaFace is used as a fallback. With `-segment face`, BiSeNet yields a binary face mask (indices $\{1, 2, 3, 4, 5, 9, 10, 11, 12, 13\}$) with hole filling; with `-segment head`, the head model is used. Default margins are 0.8 (video) and 1.0 (image). Only masked pixels are composited back into the original frame.

Attribute Extraction and Prompting

Four attributes are extracted for prompt construction: *age group*, *race*, *gender*, and *expression*. The prompt follows a fixed template: *"A photorealistic portrait of a {AGE-GROUP} {RACE} {GENDER}, with a {EXPRESSION} expression"*. A negative prompt is added to suppress anatomical distortions and cartoon-like artifacts.

Age. Numeric ages from DeepFace are mapped to coarse categories: *newborn* (< 1), *toddler* (1–3), *child* (4–9), *pre-teen* (10–13), *teenager* (14–17), *young adult* (18–24), *adult* (25–34), *middle-aged adult* (35–49), *mature adult* (50–64), *senior* (65–79), and *elderly* (≥ 80).

Race and gender. Dominant race is title-cased (e.g., *Asian*, *White*); gender is lower-cased (*female*, *male*). Empty outputs are omitted. These descriptors serve only as conditioning tokens for anonymization.

Expression. If *neutral* has confidence ≥ 0.5, it is chosen. Otherwise, the highest-probability class is selected: if confidence ≥ 0.8, it is used directly, otherwise it is softened with the prefix *mildly* (e.g., *mildly happy*). This reduces over-assertive conditioning under uncertainty.

Parsing and overrides. Custom prompts can be parsed back into the four slots to ensure consistent logging and evaluation across scenes.

Diffusion Inpainting Settings

Three ControlNets (inpaint, lineart, openpose) are employed. Inputs are resized to $\leq 612 \times 612$ and rounded to the nearest multiple of 8. Defaults are 35 denoising steps, guidance scale = 8.0, and control scales $[0.9, 0.4, 0.3]$. **Anonymity fail-safe:** if `DeepFace.verify` still matches (distance < 0.3), the system iteratively increases `strength` (+0.05), `guidance` (+1 up to 20), and `steps` (+5 up to 70), for at most three attempts.

Motion Transfer

LivePortrait (default) supports stitching (enabled by default; disable with `-no_stitch`). Per-scene outputs are merged and an additional side-by-side diagnostic video is produced. **LIA** operates at 256×256 and is well suited to short clips; the generator encodes motion and decodes frames sequentially.

Image Mode (Multiple Faces)

All detected faces are anonymized independently using shared settings and composited back. For single-face inputs, the output filename encodes the verification distance and attributes; for multi-face inputs, filenames include the face count.

Key Hyperparameters (Summary)

Component	Default / Range
Scene cut threshold	`scene_threshold` $= 0.2$
Scene RGB similarity	`scene_similarity_threshold` $= 2.0$
YOLO–ONNX	`conf` $= 0.8$, `IoU` $= 0.3$
RetinaFace margin (video / image)	0.8/1.0
Inpainting steps	$35 \rightarrow \leq 70$ (retry)
Guidance scale	$8.0 \rightarrow \leq 20$ (retry)
Control scales	$[0.9, 0.4, 0.3]$
Max input size	612×612 (rounded to /8)
Anon. verify threshold	0.3 (DeepFace distance)
Max retries	3
Motion backend	`lp` (default) or `lia`
LP stitching	`True` (disable with `-no_stitch`)

Table A.1: Key hyperparameters and default settings.

Reproducibility Notes

Identity seeds are tied to scene-level clusters to preserve cross-scene consistency. Video I/O uses OpenCV/FFmpeg (ensure `ffmpeg` is on `PATH`). Mixed precision (FP16), xFormers attention, and optional CPU offloading reduce memory pressure; we recommend $\geq$15 GB VRAM for default video settings.

A.2 Ethical Considerations and Regulatory Compliance

Face anonymization systems must be evaluated not only for utility but also for fairness and privacy robustness. A key risk is the inadvertent introduction or amplification of biases with respect to age, gender, race, or affect. While our method conditions on such attributes to preserve semantic consistency, careful auditing is required to avoid systematic skews and to document limitations.

From a regulatory perspective, frameworks such as the GDPR emphasize privacy by design and the minimization of re-identification risk. Beyond perceptual real-

ism, effective anonymization requires explicit testing against face-verification and re-identification attacks. We therefore recommend reporting privacy metrics alongside downstream-task performance, and providing clear disclosures to end users regarding residual risks.

Finally, the present system does not modify audio. Where voice privacy is a concern, integrating established audio anonymization or voice conversion components is necessary to provide end-to-end protections and to align with sector-specific compliance requirements.

Code Availability

The full implementation of *AnonNET*, including preprocessing scripts, trained models, and anonymized derivatives of VoxCeleb2, CelebV-HQ, and HDTF used in our experiments, is available at: `https://github.com/anilegin/AnonNET`.

A.3 Supplementary Results

Extended Visualizations

Figure A.1: Qualitative comparison of motion-transfer models. Each row shows consecutive frames from a video; columns correspond to the original, LivePortrait, and LIA, respectively.

Figure A.2: Examples of anonymization (part 1). For each row, the **left** image is the original first frame, the **middle** image the anonymized source, and the **right** image an anonymized video frame.

Figure A.3: Examples of anonymization (part 2). Column order matches part 1: original first frame (left), anonymized source (middle), anonymized video frame (right).