CVPR #26

EdgeDiffusion: Latency-Optimized Diffusion Models for Real-Time On-Device Visual Anonymization

Anonymous CVPR submission

Paper ID 26

Abstract

Protecting visual privacy in surveillance systems requires anonymization methods that are both effective and computationally efficient, especially when deployed on resource-constrained edge devices. Existing approaches often rely on server-side processing, exposing sensitive visual data during transmission, or degrade perceptual fidelity to meet strict latency requirements. We present ***EdgeDiffusion****, a diffusion-based framework for real-time, on-device anonymization of privacy-sensitive regions, including faces, license plates, and scene text. The framework combines three components: (i) a lightweight multi-task detector optimized for high-recall identification of privacy-sensitive regions, (ii) a distilled diffusion backbone that generates anonymized content in as few as* 4–8 *denoising steps, and (iii) category-aware anonymization modules with an adaptive fidelity controller that balances privacy protection and downstream visual utility. Anonymization effectiveness is evaluated using a Privacy Accuracy (PA) metric, which measures the reduction in successful re-identification by pretrained face, text, and license-plate recognition models after anonymization. Experiments on COCO-Privacy, Cityscapes-Privacy, and VIRAT-Privacy show that EdgeDiffusion achieves real-time inference (*$<$ 50 *ms per frame) on commercial edge hardware while improving PA by up to* 12.4% *over prior real-time baselines. EdgeDiffusion produces visually consistent anonymized outputs with limited artifacts while largely preserving scene context, suggesting its potential for privacy-preserving perception in applications such as intelligent transportation, healthcare monitoring, and smart-city analytics.*

1. Introduction

Visual data collected by surveillance systems, autonomous vehicles, and wearable devices frequently contains sensitive information such as faces, license plates, and textual identifiers. While such data supports important applications in transportation safety, healthcare monitoring, and smart-city analytics, it also introduces risks of privacy leakage, identity misuse, and regulatory non-compliance [5, 22]. Recent incidents involving unauthorized facial recognition and large-scale data breaches illustrate how easily exposed visual information can be exploited. Meanwhile, regulations such as GDPR and CCPA [3, 16, 20] impose strict requirements on the collection and processing of personal imagery, making anonymization an essential component of responsible vision systems rather than an optional safeguard [13].

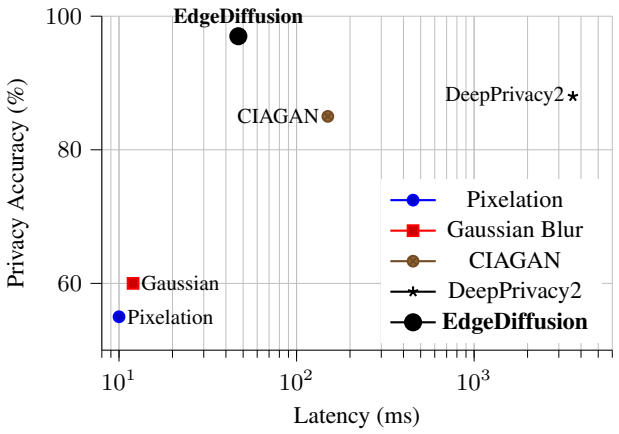

Figure 1. **Latency–Accuracy tradeoff**. EdgeDiffusion achieves superior privacy accuracy at low latency compared with baselines.

Despite substantial progress, existing anonymization approaches still struggle to simultaneously satisfy three key requirements: (*i*) effective privacy protection, reducing the risk that personally identifiable information (PII) can be reconstructed or re-identified; (*ii*) visual fidelity and utility, preserving perceptual realism while maintaining downstream analytic performance; and (*iii*) computational efficiency, enabling real-time inference on resource-constrained edge devices. Classical obfuscation techniques such as blurring and pixelation [19] are computationally efficient but often leave structural cues that remain exploitable [15]. Adversarial cloaking approaches [1, 18]

CVPR #26

can provide stronger protection under certain conditions, yet may lack robustness and can degrade downstream task performance. GAN-based anonymization methods improve perceptual realism [7, 8, 12], but their training and inference overhead limits deployment on edge devices. Diffusion models [6, 17] offer strong generative fidelity and controllability; however, conventional diffusion sampling typically requires tens to hundreds of denoising steps, making real-time or on-device deployment challenging.

To address these limitations, we introduce *EdgeDiffusion*, a unified framework for real-time, on-device visual anonymization that jointly considers privacy protection, visual fidelity, and computational efficiency. EdgeDiffusion integrates three complementary components: (i) a *privacy-oriented high-recall detector* with temporal amortization to localize diverse PII categories across video frames; (ii) a *latency-optimized diffusion backbone* distilled to operate in as few as 4–8 denoising steps while maintaining perceptual quality; and (iii) *category-aware anonymization modules* with adaptive fidelity control, enabling tailored anonymization of faces, license plates, and scene text while balancing privacy protection with downstream utility.

To quantify anonymization effectiveness, we evaluate results using a *Privacy Accuracy (PA)* metric that measures the reduction in successful re-identification by pre-trained face, text, and license-plate recognition models after anonymization. Experiments on COCO-Privacy [11], Cityscapes-Privacy [2], and VIRAT-Privacy [14] show that EdgeDiffusion achieves real-time inference on commercial edge hardware with sub-50 ms latency while maintaining consistently high anonymization effectiveness and minimal degradation in downstream detection and segmentation tasks. Compared with classical obfuscation and GAN-based approaches, EdgeDiffusion provides an improved overall trade-off between anonymization effectiveness, visual realism, and computational efficiency.

Building on this design, EdgeDiffusion forms a unified pipeline for practical on-device anonymization. Our contributions are summarized as follows:

- We propose **EdgeDiffusion**, a latency-aware diffusion framework for real-time anonymization of privacy-sensitive regions on edge devices.
- We design a **high-recall privacy detector**, a **distilled few-step diffusion backbone**, and **category-aware anonymization modules** with adaptive fidelity control, forming an integrated and practical anonymization pipeline suitable for deployment.
- We conduct **extensive experiments** on COCO-Privacy, Cityscapes-Privacy, and VIRAT-Privacy, demonstrating consistently improved privacy accuracy, strong perceptual quality, and substantially lower latency compared with existing anonymization approaches.

2. EdgeDiffusion Framework

We present **EdgeDiffusion**, a framework for *real-time, on-device* anonymization designed to reduce re-identification risk while maintaining the usability of visual data for downstream perception tasks. Given an input frame $\mathbf{x}$ and detected personally identifiable information (PII) regions $\mathcal{R} = \{r_1, \ldots, r_m\}$ such as faces, license plates, or textual identifiers, the system produces an anonymized image $\hat{\mathbf{x}}$ in which sensitive regions are modified while non-PII regions remain visually consistent with the original scene. All computation is performed under a device-specific latency constraint $T_{\max}$ to ensure feasibility in time-sensitive edge deployments. As illustrated in Figure 2, EdgeDiffusion consists of four components: (1) a high-recall multi-task detector that localizes PII regions and produces masks; (2) a latency-aware diffusion backbone that generates anonymized content using a small number of denoising steps; (3) category-aware anonymization modules that adapt the generative process to different PII categories within a shared framework; and (4) an adaptive fidelity controller that adjusts anonymization strength according to privacy–utility considerations and hardware constraints. These components together enable a pipeline for visual anonymization on resource-constrained edge devices.

2.1. Privacy-Oriented High-Recall Detection

The detection stage is based on a lightweight one-stage detector implemented on top of the YOLOv5-tiny architecture with feature pyramid fusion [9]. This design offers favorable computational efficiency on resource-constrained edge hardware while maintaining sensitivity to small-scale objects such as faces, text regions, and license plates. The detector adopts a multi-task architecture: a shared backbone $\phi(\cdot)$ (CSPDarknet-tiny with PANet feature aggregation) extracts multi-scale features $\mathbf{z}_t$ from $\mathbf{x}_t$, while three task-specific heads ψ_c are attached for $c \in \{\text{face}, \text{text}, \text{plate}\}$. The text head incorporates an EAST-style branch to better capture elongated structures, while the face and plate heads follow standard YOLO prediction layers.

Formally, given a frame $\mathbf{x}_t$, the detector $\mathcal{F}_\theta$ produces

$$\mathcal{R}_t = \mathcal{F}_\theta(\mathbf{x}_t) = \{(b_{t,i}, m_{t,i}, c_{t,i}, s_{t,i})\}_{i=1}^{M_t}, \qquad (1)$$

where $b_{t,i}$ denotes the bounding box, $m_{t,i}$ the predicted soft mask, $c_{t,i}$ the object category, $s_{t,i}$ the confidence score, and M_t the number of detected PII regions.

To bias training toward higher recall, we employ a Tversky loss with $\beta > \alpha$ that penalizes missed regions more strongly than redundant ones:

$$\mathcal{L}_{\text{tv}}(m, m^\star) = 1 - \frac{|m \cap m^\star|}{|m \cap m^\star| + \alpha |m \setminus m^\star| + \beta |m^\star \setminus m|}. \qquad (2)$$

To further reduce the risk of missed detections in challenging cases, we introduce a Conditional Value-at-Risk

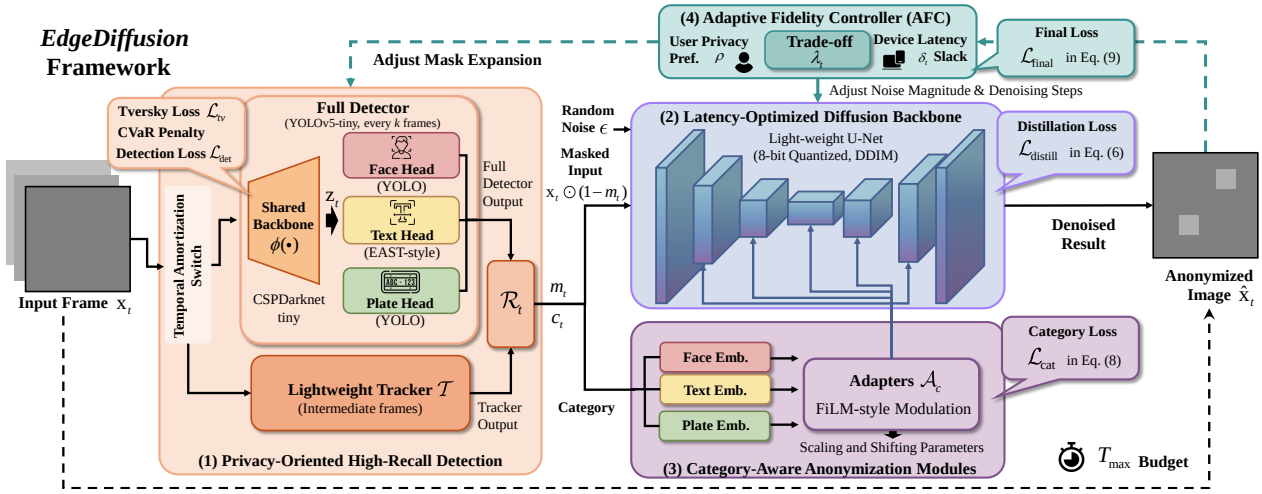

Figure 2. EdgeDiffusion framework for real-time visual anonymization. It consists of three stages: PII detection, latency-optimized diffusion-based anonymization, and category-aware control modules that regulate anonymization fidelity for different PII categories.

(CVaR) penalty that emphasizes hard samples with large detection errors. Task balancing across detection objectives is handled through uncertainty-based weighting:

$$\mathcal{L}_{\text{det}} = \sum_k \frac{1}{2\sigma_k^2}\mathcal{L}_k + \sum_k \log \sigma_k, \tag{3}$$

where $\mathcal{L}_k$ corresponds to classification, bounding box regression, and mask prediction losses.

To meet the latency constraint $T_{\max}$ on edge devices, we adopt temporal amortization. The full detector is executed every k frames, while a lightweight tracker $\mathcal{T}$ propagates the detected regions across intermediate frames:

$$\mathcal{R}_t = \begin{cases} \mathcal{F}_\theta(\mathbf{x}_t), & t \bmod k = 0, \\ \mathcal{T}(\mathcal{R}_{t-1}, \mathbf{x}_t), & \text{otherwise.} \end{cases} \tag{4}$$

This strategy reduces the average detection cost to $\frac{T_{\text{det}}}{k} + T_{\text{trk}}$. The detector is re-invoked when motion magnitude or tracking uncertainty exceeds predefined thresholds.

2.2. Latency-Optimized Diffusion Backbone

The central component of EdgeDiffusion is a diffusion-based generator that replaces sensitive regions with anonymized yet contextually consistent visual content. While diffusion models are known for strong generative fidelity, conventional implementations typically require hundreds of denoising steps, making them difficult to deploy on resource-constrained edge devices. To address this limitation, we design a latency-optimized backbone that distills the generative process into a compact few-step model while maintaining good perceptual quality.

Formally, given a masked input $\mathbf{x}_t \odot (1-m)$ and random noise $\epsilon \sim \mathcal{N}(0, I)$, a diffusion model iteratively refines a latent representation $\mathbf{z}_s$ from timestep $s = S$ down to $s = 0$:

$$\mathbf{z}_{s-1} = \frac{1}{\sqrt{\alpha_s}}\left(\mathbf{z}_s - \frac{1-\alpha_s}{\sqrt{1-\bar{\alpha}_s}}\epsilon_\theta(\mathbf{z}_s, m, s)\right) + \sigma_s \epsilon, \tag{5}$$

where ϵ_θ denotes the denoising network and $\alpha_s, \bar{\alpha}_s, \sigma_s$ are standard diffusion parameters. To enable real-time inference, we adopt a DDIM-style deterministic sampler and progressively distill the full diffusion trajectory ($S \sim 1000$) into a significantly shorter process ($S = 4-8$).

The denoising network is implemented as a lightweight U-Net with depthwise separable convolutions, while attention modules are restricted to low-resolution feature maps to control computational cost. We further prune redundant channels and apply 8-bit quantization, resulting in a $2.4\times$ reduction in memory footprint. This design keeps the backbone computation within the latency budget $T_{\max}$.

Training proceeds by aligning student predictions with a teacher diffusion model through progressive distillation:

$$\mathcal{L}_{\text{distill}} = \mathbb{E}_{\mathbf{z}_s, m}\left[\|\epsilon_\theta^{\text{student}}(\mathbf{z}_s, m, s) - \epsilon_\theta^{\text{teacher}}(\mathbf{z}_s, m, s)\|_2^2\right], \tag{6}$$

allowing the reduced-step model to closely approximate the behavior of the full diffusion process while remaining suitable for practical real-time anonymization.

2.3. Category-Aware Anonymization Modules

Different types of PII exhibit heterogeneous structures and visual contexts: faces and bodies contain fine-grained identity cues, scene text often appears as elongated high-frequency patterns, while license plates are compact but semantically important. A single anonymization strategy may therefore be less effective when applied uniformly. To address this limitation, EdgeDiffusion introduces category-aware modules that adapt the generative process to each PII type while sharing the same diffusion backbone.

CVPR #26

Concretely, the diffusion model is conditioned on both the mask m and the category label $c \in \{\text{face}, \text{text}, \text{plate}\}$. Category information is encoded as a one-hot vector and projected through a learnable embedding layer, which is injected into the U-Net backbone via lightweight adapters $\mathcal{A}_c$ inserted at multiple resolution scales (encoder, bottleneck, and decoder). These adapters are implemented as low-rank bottleneck layers with FiLM-style modulation (feature-wise linear modulation), where scaling and shifting parameters are predicted from the category embedding. Intermediate features are thus modulated as:

$$\epsilon_\theta(\mathbf{z}_s, m, s, c) = \epsilon_\theta(\mathbf{z}_s, m, s) + \mathcal{A}_c(\mathbf{z}_s, m, s), \quad (7)$$

where ϵ_θ is the shared denoising network and $\mathcal{A}_c$ provides category-specific residual adjustments. For instance, the text adapter encourages background consistency during inpainting, while the face adapter promotes natural appearance while reducing identity-related cues.

Training is performed jointly across categories. Each adapter is optimized using a combination of perceptual similarity and category-specific adversarial feedback:

$$\mathcal{L}_{\text{cat}} = \lambda_{\text{perc}} \mathcal{L}_{\text{LPIPS}} + \lambda_{\text{adv}} \, \mathbb{E}\big[\log(1 - D_c(\hat{\mathbf{x}}))\big], \quad (8)$$

where D_c is a lightweight category discriminator trained to distinguish anonymized outputs from natural samples within domain c. To stabilize training and mitigate imbalance across categories, we adopt class-balanced sampling during minibatch construction while sharing the diffusion backbone across all categories. Adapters are applied at three spatial scales (16, 32, and 64 feature maps) using depthwise separable convolutions to reduce computational overhead. Category embeddings are 64-dimensional and learned jointly with the backbone parameters. This design allows the model to adapt anonymization behavior for different PII categories without retraining the full backbone.

2.4. Adaptive Fidelity Controller

While high-recall detection and category-specific synthesis help reduce re-identification risk, practical deployments often involve heterogeneous requirements on the trade-off between anonymization strength and visual fidelity. Law-enforcement applications may prefer stronger obfuscation even at the expense of visual detail, whereas commercial analytics may favor minimal distortion to better preserve scene context. To accommodate such variability, EdgeDiffusion incorporates an *Adaptive Fidelity Controller* (AFC) that adjusts anonymization intensity according to privacy–utility preferences and device conditions.

Formally, let $\lambda \in [0, 1]$ denote the trade-off coefficient between privacy loss $\mathcal{L}_{\text{priv}}$ and utility loss $\mathcal{L}_{\text{util}}$. The overall training objective is defined as

$$\mathcal{L}_{\text{final}} = \lambda \, \mathcal{L}_{\text{priv}} + (1 - \lambda) \, \mathcal{L}_{\text{util}}. \quad (9)$$

Table 1. **Utility preservation on downstream tasks.** Relative performance drop (percentage points, ↓) in detection mAP on COCO and segmentation mIoU on Cityscapes after anonymization.

Method	COCO mAP (↓ pp)	Cityscapes mIoU (↓ pp)
No Anonymization	40.0 (0.0)	74.0 (0.0)
Pixelation	34.5 (5.5)	68.5 (5.5)
Gaussian Blur	35.5 (4.5)	69.5 (4.5)
CIAGAN	37.8 (2.2)	71.2 (2.8)
DeepPrivacy2	38.3 (1.7)	71.6 (2.4)
EdgeDiffusion	**38.8 (1.2)**	**72.0 (2.0)**

Where $\mathcal{L}_{\text{priv}}$ penalizes residual identity cues through auxiliary re-identification classifiers and embedding similarity constraints, while $\mathcal{L}_{\text{util}}$ aims to preserve downstream task performance and perceptual quality, measured using standard metrics such as SSIM, LPIPS, and detection mAP.

At inference time, λ is adjusted according to (i) a user-specified privacy preference $\rho \in [0, 1]$, and (ii) the device's latency slack $\delta_t = T_{\max} - T_{\text{curr}}$. We define

$$\lambda_t = \sigma\big(\alpha\rho + \beta\delta_t\big), \quad (10)$$

where $\sigma(\cdot)$ is a sigmoid and (α, β) determine the relative influence of application-level priorities and hardware state. A higher ρ typically corresponds to stronger anonymization preference, while positive δ_t allows slightly additional computation that may improve visual fidelity.

In practice, AFC modulates the diffusion process through two mechanisms: (1) adjusting the injected noise magnitude and the effective number of denoising steps; and (2) varying mask expansion and backbone compression depending on current device load. Under stronger privacy preference, AFC increases noise variance and mask dilation; under relaxed conditions, it reduces noise injection to better preserve fine-grained textures. This mechanism allows the system to adapt anonymization behavior to different application requirements and hardware constraints.

3. Experiments

3.1. Experimental Setup

We evaluate **EdgeDiffusion** under deployment-oriented conditions reflecting its intended use case: real-time anonymization on commercial edge platforms. The evaluation protocol considers three complementary aspects: (1) *privacy protection*, assessing the system's ability to reduce identifiable information such as faces, license plates, and scene text; (2) *utility preservation*, measuring the extent to which anonymized imagery continues to support downstream perception tasks (e.g., detection and segmentation); and (3) *efficiency*, quantifying end-to-end latency and memory footprint on resource-constrained hardware.

Experiments are conducted on three complementary benchmarks: COCO-Privacy [11], Cityscapes-Privacy [2],

Figure 3. **Examples of face and full-body anonymization.** EdgeDiffusion preserves background and clothing structures while anonymizing facial and body regions that may contain personally identifiable information. Compared with pixelation, blurring, and GAN-based approaches, EdgeDiffusion generally provides improved visual consistency while maintaining anonymization.

Table 2. **Face/Full-Body anonymization on COCO-Privacy.** PA denotes Privacy Accuracy (higher is better), measuring the reduction of successful re-identification after anonymization.

Method	PA (↑ %)	SSIM (↑)	LPIPS (↓)	Detection mAP (↑ %)	Latency (↓ ms)	Memory (↓ MB)
Pixelation	55.0	0.81	0.21	52.0	10	80
Gaussian Blur	60.0	0.84	0.19	55.0	12	80
CIAGAN (CVPR 2020)	85.0	0.90	0.15	62.0	150	420
DeepPrivacy2 (WACV 2023)	88.0	0.92	0.13	64.0	3600	550
EdgeDiffusion (Face/Full-Body)	**97.0**	**0.94**	**0.10**	**66.0**	**47**	**310**

and VIRAT-Privacy [14], each annotated with privacy-sensitive regions. We adopt a consistent 70/10/20 (train/val/test) split with category-balanced sampling to facilitate reproducible comparison.

For deployment evaluation, we target two representative hardware platforms: *NVIDIA Jetson Orin Nano*, an embedded GPU platform widely used in robotics and autonomous systems [4], and *Qualcomm Snapdragon 8 Gen 2*, a modern mobile SoC representative of consumer devices. Performance metrics are reported as average wall-clock latency and peak memory usage measured directly on-device over 1,000 test frames at 640×480 resolution.

3.2. Datasets

To evaluate **EdgeDiffusion**, we benchmark on three datasets that capture privacy-sensitive scenarios spanning a range of PII types and environmental conditions. 1) **COCO-Privacy [11]** extends Microsoft COCO [11] with

Table 3. **Scene text removal performance on COCO-Privacy.** Detection mAP Δ denotes the performance drop (percentage points) relative to the original images; lower values indicate better preservation of downstream detection performance.

Method	PSNR (↑)	SSIM (↑)	LPIPS (↓)	Detection mAP Δ (pp) (↓)	Latency (↓ ms)	Memory (↓ MB)
Baseline (EAST + Inpainting)	30.0	0.85	0.20	10.0	80	100
GaRNet (ECCV 2022) [10]	32.5	0.88	0.17	7.0	200	300
PERT (WACV 2023) [21]	33.5	0.90	0.15	5.0	250	350
EdgeDiffusion (Text)	**35.0**	**0.92**	**0.12**	**3.0**	**50**	**310**

Table 4. **License plate anonymization on Cityscapes-Privacy.** LP readability denotes the recognition accuracy of license plates after anonymization (lower is better). Detection mAP Δ indicates the performance drop (percentage points) relative to the original images.

Method	LP Readability (↓ %)	LPIPS (↓)	Detection mAP Δ (pp) (↓)	Latency (↓ ms)
Blur (Gaussian)	75.0	0.18	8.0	10
dashcam_anonymizer	60.0	0.15	5.0	120
EdgeDiffusion (Plate)	**45.0**	**0.10**	**3.0**	**47**

annotations of privacy-sensitive regions, including human faces, full-body regions, vehicle license plates, and scene text. The diversity of scenes, cluttered environments, and occlusions provides a challenging benchmark for multi-category anonymization. 2) **Cityscapes-Privacy [2]** focuses on license plate and pedestrian identity protection in urban driving environments. It provides pixel-level masks for vehicles, license plates, and pedestrians, enabling evaluation of anonymization quality as well as downstream perception tasks. 3) **VIRAT-Privacy [14]** provides temporal annotations for faces, vehicles, and textual identifiers across surveillance videos. The dataset includes dynamic scenes with varying illumination and crowd activity, supporting evaluation of temporal consistency in streaming settings.

3.3. Face and Full-Body Anonymization

Table 2 compares **EdgeDiffusion** with classical obfuscation methods (Pixelation, Gaussian Blur) and representative GAN-based anonymization approaches (CIAGAN [12], DeepPrivacy2 [7]) on the COCO-Privacy dataset. Classical obfuscation methods achieve negligible latency but provide limited privacy protection, with Privacy Accuracy (PA) below 60% and residual identity cues that can still be recognized by standard recognition models. GAN-based approaches improve PA (85–88%) and perceptual realism, but incur substantially higher latency (150 ms for CIAGAN and >3.6 s for DeepPrivacy2) together with larger memory footprints, which can limit their applicability on edge platforms. In contrast, **EdgeDiffusion** achieves higher anonymization effectiveness, reaching 97% PA while maintaining high structural similarity (SSIM = 0.94) and low perceptual distance (LPIPS = 0.10). The method preserves downstream task performance while operating at 47 ms latency with a 310 MB memory footprint. Qualitative results in Figure 3 illustrate that EdgeDiffusion generates anonymizations that blend naturally with surrounding context.

3.4. Scene Text Removal

Scene text anonymization is challenging due to elongated, high-frequency patterns that can introduce visible artifacts when removed. Table 3 compares **EdgeDiffusion** with inpainting-based methods (EAST+Inpainting), GaRNet [10], and PERT [21]. EdgeDiffusion achieves higher reconstruction fidelity (PSNR = 35.0, SSIM = 0.92, LPIPS = 0.12) while maintaining smaller detection mAP degradation (↓3.0 pp) than the baselines (↓5–10 pp). Its inference latency (50 ms) is also lower than GaRNet (200 ms) and PERT (250 ms), with a memory footprint comparable to other anonymization settings. Figure 4 further illustrates that the method removes text regions while preserving natural background textures and maintaining visual consistency.

3.5. License Plate Anonymization

Table 4 evaluates license plate anonymization on the Cityscapes-Privacy benchmark. Classical pixelation-based blur provides limited anonymization, with license plate (LP) readability remaining as high as 75% together with noticeable visual distortion. The dashcam_anonymizer baseline reduces readability to 60% but incurs higher latency (120 ms), which can limit real-time deployment. In comparison, **EdgeDiffusion** achieves lower LP readability (**45%**) while maintaining relatively low perceptual distortion (LPIPS = 0.10) and smaller detection accuracy degradation (3 pp), operating at 47 ms per frame. Qualitative examples in Fig. 4 further illustrate that **EdgeDiffusion** removes alphanumeric identifiers while preserving plate geometry and surrounding structures, producing overall visually consistent anonymization results.

3.6. End-to-End Efficiency

Table 5 compares system efficiency across heterogeneous edge platforms. GAN-based anonymization methods such

Figure 4. **Examples of scene text anonymization.** EdgeDiffusion modifies sensitive textual and license plate regions while preserving surrounding structures and visual continuity. The examples illustrate anonymization results on commercial signage and retail wall text.

Table 5. **End-to-end efficiency on edge devices** (640×480 resolution). Latency, throughput (FPS), and peak memory usage are reported for two representative edge platforms. Best results are shown in bold.

Method	Jetson Orin Nano			Snapdragon 8 Gen 2		
	Latency (↓ ms)	FPS (↑)	Peak Memory (↓ MB)	Latency (↓ ms)	FPS (↑)	Peak Memory (↓ MB)
CIAGAN	150	6.7	420	180	5.6	430
DeepPrivacy2	3600	0.3	550	4000	0.25	600
GaRNet	200	5.0	300	240	4.2	320
PERT	250	4.0	350	300	3.3	370
dashcam_anonymizer	120	8.3	300	140	7.1	320
EdgeDiffusion	**47**	**21.3**	**310**	**55**	**18.2**	**320**

as CIAGAN and DeepPrivacy2 exhibit high runtime and memory requirements (6.7 FPS at 420 MB; 0.3 FPS at 550 MB), which can limit real-time deployment. Scene text removal networks (GaRNet, PERT) incur higher computational costs (200–300 ms per frame), further constraining deployment on resource-limited devices. In contrast, **EdgeDiffusion** achieves **21.3 FPS** on Jetson Orin Nano and **18.2 FPS** on Snapdragon 8 Gen 2, with a memory footprint of 310–320 MB. These results indicate that the latency-optimized diffusion backbone (Sec. 2.2) operates efficiently across both embedded GPU and mobile SoC platforms.

The latency–accuracy trade-off illustrated in Fig. 1 further shows that EdgeDiffusion achieves **97% Privacy Accuracy** at a real-time latency of 47 ms. Classical obfuscation methods remain faster but provide limited anonymization, while GAN-based approaches offer stronger protection at substantially higher computational cost.

3.7. Cross-Dataset Generalization

Table 7 evaluates cross-dataset generalization by training on COCO-Privacy and testing on four external datasets: LFW, WiderFace, BDD100K, and Cityscapes. GAN-based baselines retain moderate Privacy Accuracy (80–87%) but exhibit noticeable degradation in perceptual fidelity and downstream mAP when transferred across domains. In comparison, **EdgeDiffusion** achieves **91–95% PA**, with low perceptual distortion (LPIPS = 0.11–0.12) and stable task performance (mAP = 64–65%). These results suggest that the method maintains consistent performance across different datasets and visual conditions.

3.8. Utility Preservation

Table 1 evaluates anonymization as a preprocessing stage for downstream perception tasks, including object detection (COCO mAP) and semantic segmentation (Cityscapes mIoU). Classical obfuscation methods introduce noticeable utility degradation, causing 5–6 pp drops in accuracy. GAN-based approaches preserve more task performance but still exhibit larger degradation relative to the non-anonymized baseline. In comparison, our **EdgeDiffusion** maintains higher downstream performance, reducing the detection accuracy by only **1.2 pp** and the segmentation accuracy by **2.0 pp**. Therefore, these results indicate that EdgeDiffusion preserves compatibility with downstream perception tasks while performing anonymization.

Table 6. **Ablation Study on COCO-Privacy Test Set.** Latency comparisons to the full version. Best results are in bold.

Variant	PA (↑ %)	SSIM (↑)	LPIPS (↓)	Latency (↓ ms)
Full EdgeDiffusion	**97.0**	**0.94**	**0.098**	**47**
w/o Adaptive Fidelity Control	93.0	0.92	0.11	47
w/o Optimized Diffusion Backbone	96.0	0.94	0.10	260
w/o Uncertainty-Aware Mask Expansion	92.0	0.93	0.11	47

Table 7. **Cross-dataset generalization.** Train on COCO-Privacy, test on multiple datasets. Best results are in bold.

	LFW				WiderFace				BDD100K				Cityscapes			
Method	**PA**	**SSIM**	**LPIPS**	**mAP**	**PA**	**SSIM**	**LPIPS**	**mAP**	**PA**	**SSIM**	**LPIPS**	**mAP**	**PA**	**SSIM**	**LPIPS**	**mAP**
Pixelation	56.0	0.82	0.21	52.5	54.0	0.81	0.22	51.0	53.0	0.83	0.21	51.5	50.0	0.82	0.22	50.0
Gaussian	61.0	0.85	0.19	55.5	59.0	0.84	0.20	54.0	58.0	0.86	0.19	54.5	55.0	0.85	0.20	53.0
CIAGAN	83.5	0.90	0.15	61.5	82.0	0.89	0.16	60.5	80.0	0.90	0.15	60.0	78.0	0.89	0.16	59.0
DeepPrivacy2	86.5	0.92	0.13	63.0	85.0	0.91	0.14	62.0	83.0	0.92	0.13	62.0	82.0	0.91	0.14	61.0
EdgeDiffusion	**95.0**	**0.93**	**0.11**	**65.0**	**94.0**	**0.92**	**0.12**	**64.0**	**92.0**	**0.93**	**0.11**	**65.0**	**91.0**	**0.92**	**0.12**	**64.0**

Table 8. **Module-wise latency breakdown** on Jetson Orin Nano.

Component	Latency (ms)	Share (%)
Detector (amortized)	8	17
Tracker	3	6
Mask Expansion	2	4
Diffusion Backbone	31	66
Post-processing	3	6
Total	**47**	**100**

3.9. Ablation Studies

We analyze the contribution of each component through controlled ablations on COCO-Privacy (Table 6). Disabling *Adaptive Fidelity Control* reduces Privacy Accuracy (PA) from 97% to 93% and increases LPIPS, indicating its role in balancing anonymization strength and perceptual quality. Replacing the optimized diffusion backbone with a standard U-Net increases latency from 47 ms to 260 ms while providing limited accuracy improvement, highlighting the efficiency benefit of the distilled backbone. Finally, removing *uncertainty-aware mask expansion* decreases PA by 5 pp, leaving residual PII around mask boundaries. Qualitative results in Fig. 4 further illustrate the contribution of each component to the overall anonymization performance.

3.10. Runtime Breakdown

We further profile the module-wise latency of **EdgeDiffusion** on Jetson Orin Nano (Table 8). The diffusion backbone accounts for the majority of runtime (66%), reflecting the computational cost of generative inference. Detection (17%) and tracking (6%) are amortized across frames, while mask expansion and post-processing introduce relatively small overhead. The total latency remains 47 ms per frame, with computation distributed across the major system modules, indicating that the pipeline is suitable for practical deployment on edge devices.

3.11. Limitations

While EdgeDiffusion performs well in the evaluated settings, several limitations should be noted. First, all experiments are conducted at 640×480 resolution; scaling to higher resolutions may require additional architectural adaptations, such as patch-level parallel diffusion or cascaded generation, to maintain real-time throughput on edge hardware. Second, the robustness of EdgeDiffusion against adversarial reconstruction attacks, in which an attacker attempts to reverse anonymization and recover original identities, has not been evaluated. Investigating resilience to such attacks is an important direction for future work, particularly for deployment in security-sensitive scenarios.

4. Conclusion

In this paper, we presented **EdgeDiffusion**, a diffusion-based anonymization framework designed for real-time deployment on resource-constrained edge devices, integrating a high-recall multi-task PII detector with temporal amortization, a progressively distilled few-step diffusion backbone, and category-aware anonymization modules governed by an adaptive fidelity controller. Experiments across three benchmarks and two edge platforms confirm that EdgeDiffusion achieves sub-50 ms latency with over 95% Privacy Accuracy while limiting downstream task degradation to within 2 pp, with consistent generalization across diverse visual domains. Current limitations include evaluation restricted to 640×480 resolution and the absence of robustness analysis against adversarial reconstruction attacks. Future work will explore higher-resolution multi-camera streaming with explicit temporal consistency constraints and joint visual-audio anonymization.

References

[1] Valeriia Cherepanova, Micah Goldblum, Matthew Fredrikson, and Tom Goldstein. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. In *International Conference on Learning Representations (ICLR)*, 2021. 1

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. Cityscapes benchmark. 2, 4, 6

[3] Anup Kumar Das et al. European union's general data protectionregulation, 2018: A brief overview. *Annals of Library and Information Studies (ALIS)*, 65(2):139–140, 2018. 1

[4] Yifei Dong, Fengyi Wu, Qi He, Zhi-Qi Cheng, Heng Li, Minghan Li, Zebang Cheng, Yuxuan Zhou, Jingdong Sun, Qi Dai, et al. Ha-vln 2.0: An open benchmark and leaderboard for human-aware navigation in discrete and continuous environments with dynamic multi-human interactions. *arXiv preprint arXiv:2503.14229*, 2025. 5

[5] Yifei Dong, Fengyi Wu, Sanjian Zhang, Guangyu Chen, Yuzhi Hu, Masumi Yano, Jingdong Sun, Siyu Huang, Feng Liu, Qi Dai, et al. Securing the skies: A comprehensive survey on anti-uav methods, benchmarking, and future directions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6659–6673, 2025. 1

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[7] 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. 2, 6

[8] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In *Advances in Visual Computing*, pages 565–578, Cham, 2019. Springer International Publishing. 2

[9] Rahima Khanam and Muhammad Hussain. What is yolov5: A deep look into the internal features of the popular object detector. *arXiv preprint arXiv:2407.20892*, 2024. 2

[10] Hyeonsu Lee and Chankyu Choi. The surprisingly straightforward scene text removal method with gated attention and region of interest generation. In *European Conference on Computer Vision (ECCV)*, pages 457–472. Springer, 2022. 6

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 4, 5

[12] Max Maximov, Ismail Elezi, and Laura Leal-Taixé. Ciagan: Conditional identity anonymization generative adversarial network. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 6

[13] Blaž Meden, Peter Rot, Philipp Terhörst, et al. Privacy–enhancing face biometrics: A comprehensive survey. *IEEE Transactions on Information Forensics and Security*, 16: 4147–4183, 2021. 1

[14] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, J.K̃. Aggarwal, Hyungtae Lee, Larry Davis, Eran Swears, Xiaoyang Wang, Qiang Ji, Krishna Reddy, Mubarak Shah, Carl Vondrick, Hamed Pirsiavash, Deva Ramanan, Jenny Yuen, Antonio Torralba, Bi Song, Amit Roy-Chowdhury, and Mita Desai. A large-scale benchmark dataset for event recognition in surveillance video. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3153–3160, 2011. VIRAT Video Dataset. 2, 5, 6

[15] Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3706–3715, 2018. 1

[16] Stuart L Pardau. The california consumer privacy act: Towards a european-style privacy regime in the united states. *J. Tech. L. & Pol'y*, 23:68, 2018. 1

[17] 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*, pages 10684–10695, 2022. 2

[18] Shawn Shan, Emily Wenger, Jiayun Zhang, Huan Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *USENIX Security Symposium*, 2020. 1

[19] Warit Sirichotedumrong, Yuma Kinoshita, and Hitoshi Kiya. Pixel-based image encryption without key management for privacy-preserving deep neural networks. *IEEE Access*, 7: 177844–177855, 2019. 1

[20] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676), 2017. 1

[21] Yuxin Wang, Hongtao Xie, Shancheng Fang, Yadong Qu, and Yongdong Zhang. Pert: A progressively region-based network for scene text removal, 2021. 6

[22] Ruoyu Zhao, Yushu Zhang, Tao Wang, Wenying Wen, Yong Xiang, and Xiaochun Cao. Visual content privacy protection: A survey. *ACM Computing Surveys*, 57(5):1–36, 2025. 1