GENERATIVE VIDEO ANONYMIZATION: WHOLE-BODY IDENTITY SUPPRESSION WITH SEMANTIC PRESERVATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern person re-identification systems can recover identity from whole-body biometric cues such as gait, body shape, clothing configuration, and silhouette, making face-only anonymization insufficient for privacy-preserving video analysis. Existing anonymization methods either destroy semantic utility through coarse obfuscation or preserve full-body identity cues by operating only on facial regions. We introduce *Generative Video Anonymization* (GVA), a new task that aims to synthesize a semantically equivalent video in which whole-body biometric identity is suppressed while action dynamics, scene context, and visual realism are preserved. We first formalize GVA as a constrained identity-action decoupling problem and show that general-purpose video-to-video editing methods are structurally misaligned with this objective, failing due to inadequate subject grounding, the absence of Re-ID-aware identity objectives, and weak temporal identity consistency. To support principled evaluation, we propose a dedicated privacy-utility protocol with metrics for assessing identity retrieval resilience, feature-level identity leakage, scene preservation, temporal coherence, and behavioral action consistency. We further introduce the *Belvedere Museum* dataset, a naturalistic egocentric benchmark collected in a real museum environment, where visitors and bystanders must be distinguished from artistic depictions such as paintings and statues. We present *MOSAIC*, a task-specific framework that combines agentic perceptual grounding, counterfactual identity planning, and mask-constrained diffusion-based video generation. Experiments on OpenHumanVid and Belvedere Museum show that MOSAIC achieves the strongest whole-body identity suppression among the evaluated generative methods across all three privacy metrics, while retaining competitive, metric-dependent utility, occupying a distinct operating point on the privacy–utility trade-off.

1 INTRODUCTION

Video data is increasingly central to perception, analytics, and human behavior understanding, yet its release remains constrained by privacy risks. A key challenge is that identity in video is not localized to the face. Modern Person Re-Identification (Re-ID) systems exploit distributed whole-body biometric cues, including gait, body shape, silhouette, and clothing layout, to match individuals across non-overlapping camera views He et al. (2021); Ye et al. (2021); Chao et al. (2019); Gu et al. (2022). Consequently, face-only anonymization is no longer sufficient, as even when faces are fully occluded, Re-ID models can retain substantial identification accuracy, thus exposing a gap between common anonymization practice and meaningful protection against modern biometric adversaries Oh et al. (2016).

Existing anonymization paradigms do not resolve this gap. Classical obfuscation methods, such as blurring or masking, can suppress identity cues but often destroy the semantic information that makes video useful for downstream tasks such as action recognition Ryoo et al. (2017). Face-centric generative anonymization methods produce more realistic outputs Maximov et al. (2020); Hukkelås et al. (2019); Hukkelås & Lindseth (2023), but they primarily modify facial regions and therefore leave full-body Re-ID signatures intact. These limitations reveal a broader problem as video anonymization

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

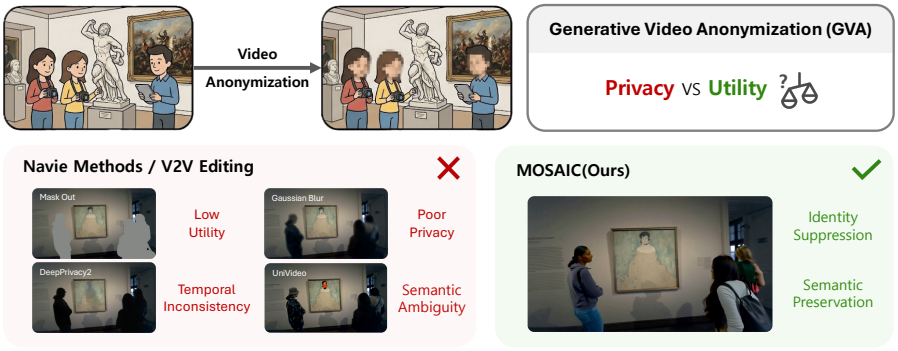

Figure 1: **The Generative Video Anonymization (GVA) Task.** (Top) GVA jointly targets biometric privacy and semantic utility. (Bottom-left) Naive methods and V2V editing fail in characteristic ways. (Bottom-right) MOSAIC suppresses whole-body identity while preserving scene and action.

should not merely hide pixels or replace faces, but should decouple biometric identity from the action dynamics and scene semantics that must remain usable.

This decoupling perspective turns anonymization into a fundamentally generative problem. An anonymized video should differ from the original in identity-bearing appearance, while remaining equivalent in action, spatial context, and scene content. This makes standard perceptual notions of fidelity insufficient: penalizing visual deviation from the original can directly conflict with the privacy objective, since successful anonymization must move the subject away from the original identity in Re-ID feature space. We therefore introduce **Generative Video Anonymization (GVA)**, the task of synthesizing a semantically equivalent video in which whole-body biometric identity is suppressed while action dynamics, scene context, and visual realism are preserved. We formalize GVA as a constrained identity-action decoupling problem in Section 2.

Recent recent video-to-video (V2V) editing methods Wu et al. (2023); Geyer et al. (2024); Wei et al. (2026) struggle with this challenge. Currently employed V2V editing models can induce visible appearance changes, but they are not designed to optimize identity displacement under semantic constraints. Our analysis identifies three structural mismatches with GVA: *region awareness*, since prompt-driven edits may alter background or non-target content; *identity awareness*, since visual change does not necessarily reduce Re-ID similarity; and *temporal identity consistency*, since weakly anchored edits can produce unstable synthetic identities over time. These failures motivate a task-specific formulation, evaluation protocol, and baseline for whole-body identity suppression.

Our contributions are summarized as follows. C1 defines the problem, C2–C3 make it measurable, and C4 provides the methodological contribution.

(C1) Task formulation and diagnostic analysis. We formulate **Generative Video Anonymization** as a constrained generative task of whole-body identity-action decoupling. We further analyze why general-purpose V2V editing methods are structurally misaligned with this objective, both conceptually and empirically (Section 2).

(C2) Evaluation protocol. We propose a dedicated protocol to formally evaluate the GVA privacy-utility trade-off. Our framework assesses identity obfuscation (via Rank-1, mAP, and $\mathcal{S}_{cos}$) alongside video utility (via SPS, TWE, and $\mathcal{A}_{cons}$), providing a holistic measure of retrieval resilience, temporal coherence, and semantic retention (Section 4.2).

(C3) Benchmark dataset. We introduce the **Belvedere Museum dataset**, an ego-centric benchmark captured with the scene camera of wearable Tobii Pro eye-tracking glasses in a real museum environment. The data were collected during regular museum visits with tourists moving naturally through the exhibition space, without a special or controlled recording setup. Beyond motion blur, viewpoint changes, and challenging illumination, this dataset requires distinguishing biological subjects from artistic depictions such as paintings and statues, a semantic ambiguity that is central to real-world privacy filtering (Section 4).

(C4) Task-specific baseline. We present **MOSAIC** (Mask-Oriented Synthesis via Agentic Identity Counterfactuals), a baseline framework for GVA. MOSAIC combines agentic perceptual grounding for subject-level privacy masks, counterfactual identity planning for identity-distant generation targets,

and mask-constrained diffusion-based video generation for temporally coherent identity replacement (Section 3).

Experiments on **OpenHumanVid** Li et al. (2025) and **Belvedere Museum** show that MOSAIC achieves the strongest identity suppression among the evaluated generative methods, while occupying a different privacy–utility operating point from existing generative baselines rather than dominating them on every metric.

2 Problem Formulation

We formalize *Generative Video Anonymization* (GVA) as a constrained identity-transformation problem. Let $V = \{I_t\}_{t=1}^{T} \in \mathbb{R}^{T\times H\times W\times 3}$ denote an input video of T frames containing identifiable subjects. GVA seeks to learn a transformation $\mathcal{T}$ that generates an anonymized counterpart $\tilde{V} = \mathcal{T}(V)$. A successful transformation must satisfy three simultaneous conceptual objectives: (i) suppress whole-body biometric identity, (ii) preserve behavioral semantics and scene context, and (iii) maintain temporal realism. Formally, the optimal transformation $\mathcal{T}^*$ minimizes identity leakage while satisfying strict utility constraints:

$$\mathcal{T}^* = \arg\min_{\mathcal{T}} \mathcal{L}_{id}\big(V, \mathcal{T}(V)\big) \quad \text{s.t.} \quad \mathcal{C}_{act}\big(\mathcal{T}(V), V\big) \geq \gamma,\ \mathcal{C}_{bg}\big(\mathcal{T}(V), V\big) \geq \eta,\ \mathcal{C}_{temp}\big(\mathcal{T}(V)\big) \geq \tau \tag{1}$$

Here, the objective function $\mathcal{L}_{id}$ conceptualizes the identity leakage within a joint biometric feature space. The constraint functions $\mathcal{C}_{act}$, $\mathcal{C}_{bg}$, and $\mathcal{C}_{temp}$ mathematically represent the preservation of action semantics, background integrity, and temporal coherence, respectively, with γ, η, and τ acting as their theoretical acceptability thresholds. *(We instantiate these abstract concepts with concrete evaluation metrics in Section 4.2.)*

This formulation reveals a critical insight: **the optimization target and the utility constraints operate in disjoint feature spaces.** Identity suppression requires traversal within a biometric manifold, whereas utility is strictly defined by spatial-temporal dynamics and visual stability. Consequently, standard generative metrics computed directly against the source video V are **not** valid proxies for GVA quality. Effective anonymization inherently forces $\mathcal{T}(V)$ to visually deviate from V to break biometric signatures; penalizing this necessary deviation conflates privacy protection with a loss of utility.

2.1 Why Video Editing Methods Fail

When applied to the GVA task, state-of-the-art Video-to-Video (V2V) editing methods (e.g., Uni-Video Wei et al. (2026)) exhibit systematic failures. These shortcomings stem from structural misalignments between general generative objectives and the strict GVA constraints defined above. We categorize these failures into three conceptual gaps that directly motivate the design of our framework, MOSAIC:

F1: Spatial Grounding Gap. V2V transformations typically apply prompt-guided modifications globally across the frame. Without precise, subject-level spatial grounding, semantic edits bleed indiscriminately into the background. This leads to severe environmental distortion and hallucinations, violating the theoretical constraint ($\mathcal{C}_{bg}$) for strict scene preservation.

F2: Identity Objective Gap. V2V models optimize for perceptual alignment with a text prompt rather than biometric displacement. While they often produce visible pixel-level alterations (e.g., changing clothing style or color), these superficial edits generally fail to deceive robust feature extractors. Consequently, outputs retain significant biometric leakage ($\mathcal{L}_{id}$), proving that superficial visual editing does not equate to robust identity suppression in a Re-ID manifold.

F3: Temporal Coherence Gap. Traditional V2V techniques lack explicit cross-frame feature anchoring for the *newly synthesized* subject. Without constraints to enforce identity consistency over time, the synthetic subject exhibits a "drifting" appearance and flickering artifacts across the

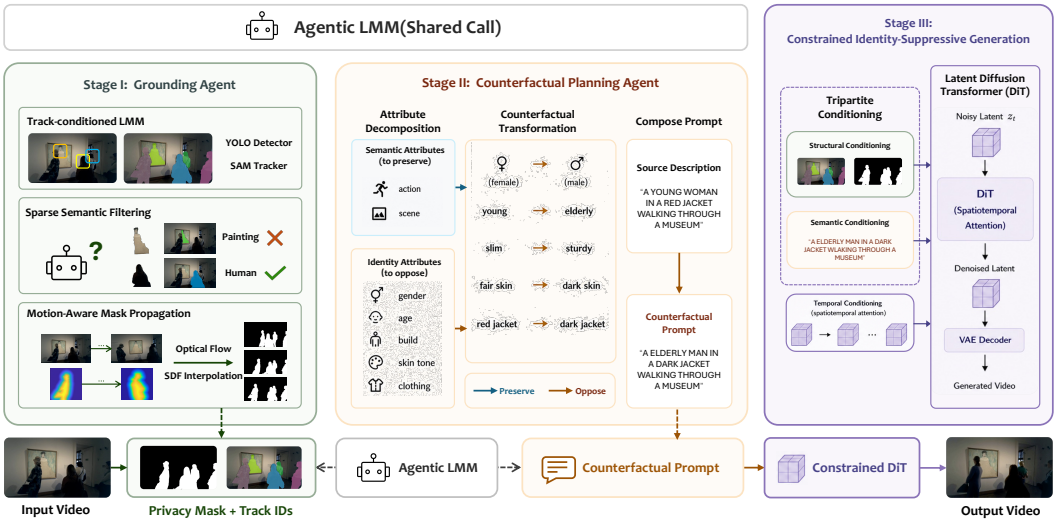

Figure 2: **Overview of MOSAIC.** Our pipeline consists of three agentic stages: (1) *Grounding Agent* for precise subject segmentation and track-based filtering; (2) *Counterfactual Planning Agent* for synthesizing identity-distant prompts via attribute decomposition; and (3) *Constrained Generation* using a DiT with tripartite conditioning to ensure spatio-temporal consistency.

sequence. This fails to establish a consistent, persistent persona, violating the temporal realism ($\mathcal{C}_{temp}$) inherently required by video anonymization.

3 MOSAIC: Mask-Oriented Synthesis via Agentic Identity Counterfactuals

We present **MOSAIC**, a three-stage framework for GVA that directly addresses the three structural gaps identified in Section 2. As illustrated in Figure 2, MOSAIC operates on an input video V through: a *Grounding Agent* that efficiently locates and tracks privacy-sensitive subjects while resolving semantic ambiguities (Section 3.1); a *Counterfactual Planning Agent* that constructs a structured identity-changing generation target (Section 3.2); and a *Constrained Generative Backbone* that produces temporally coherent, identity-suppressed video conditioned on the mask and counterfactual prompt (Section 3.3). The first two stages are unified under an agentic design that shares the same LMM (Large Multimodal Models) call, reducing computational overhead while enabling joint grounding and planning.

3.1 Stage I: Grounding Agent

The grounding agent addresses F1 by producing precise per-frame privacy masks that correctly isolate biological subjects from non-biometric depictions. It operates in two steps: sparse semantic filtering and motion-aware mask propagation.

Sparse Semantic Filtering. Applying an LMM to every frame is computationally prohibitive and redundant for subjects that persist across frames. We adopt a track-conditioned inference strategy built on a multi-object tracker that assigns a persistent track ID to each detected candidate. The LMM Carion et al. (2026); Wang et al. (2024a); Bai et al. (2023) is invoked only when a new track ID is first observed, and its output is cached for the lifetime of that track. This reduces LMM calls from $\mathcal{O}(T)$ per-frame invocations to $\mathcal{O}(N_{\text{tracks}})$ per-track invocations, a reduction of over 90% on typical video sequences (Section 4).

For each newly observed track, we construct a visual context tuple $\mathbf{x}_t^{(i)} = (\mathcal{C}(\mathbf{V}_t, \mathbf{m}_t^{(i)}), \mathbf{V}_t)$ comprising the local crop $\mathcal{C}(\cdot)$ and global frame $\mathbf{V}_t$, and query the LMM to estimate the posterior probability of the privacy class y_{priv}:

$$p_t^{(i)} = P(y = y_{\text{priv}} \mid \mathbf{x}_t^{(i)}; \Theta) \tag{2}$$

The verified mask set is obtained via a hard decision boundary τ: $\hat{\mathcal{M}}_t = \{\mathbf{m}_t^{(i)} \in \mathcal{M}_t \mid p_t^{(i)} > \tau\}$. Probabilistic filtering suppresses semantic false positives such as statues and paintings while maintaining high recall for biological subjects, directly addressing the museum disambiguation challenge of the Belvedere dataset.

Motion-Aware Mask Propagation. Verified masks are propagated to subsequent frames using a dual-mechanism motion model. For continuously tracked subjects, we enforce temporal consistency via dense optical flow. Let $\mathbf{u}_{t-1\to t} \in \mathbb{R}^{H\times W\times 2}$ be the flow field estimated by the Farneback algorithm Farnebäck (2003). The warping operator $\mathcal{W}(\cdot, \mathbf{u})$ propagates the mask forward:

$$\tilde{\mathbf{m}}_t = \mathcal{W}(\mathbf{m}_{t-1},\, \mathbf{u}_{t-1\to t}) \tag{3}$$

where $\mathcal{W}$ performs bilinear sampling at flow-displaced coordinates. For occlusion intervals $[t_a, t_b]$ where detection is unreliable, we interpolate in the Signed Distance Field (SDF) domain Osher & Sethian (1988) rather than blending pixels directly, preserving geometric topology. Let $\phi_t : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$ be the level-set representation of a mask such that $\mathbf{m}_t(\mathbf{x}) = 1 \iff \phi_t(\mathbf{x}) < 0$. The interpolated geometry at intermediate time t is:

$$\phi_t(\mathbf{x}) = (1-\alpha)\,\phi_{t_a}(\mathbf{x} - \Delta\mathbf{u}_a) + \alpha\,\phi_{t_b}(\mathbf{x} - \Delta\mathbf{u}_b) \tag{4}$$

where $\alpha = (t - t_a)/(t_b - t_a)$, and Δu_a and Δu_b denote the cumulative optical-flow displacements from the endpoint frames t_a and t_b to the intermediate frame t. Thus $x - \Delta u_a$ and $x - \Delta u_b$ are the motion-compensated coordinates at which the endpoint SDFs ϕ_{t_a} and ϕ_{t_b} are evaluated, and α interpolates them according to the temporal position of t.

3.2 Stage II: Counterfactual Planning Agent

The planning agent addresses F2 by constructing a generation prompt that explicitly specifies identity attributes differing from those of the source subject. It is executed once per track at the same LMM Wang et al. (2024a); Bai et al. (2023) call used for semantic filtering, adding negligible overhead to Stage I. Standard video captioners produce generic descriptions that preserve action semantics but provide no mechanism for identity displacement. MOSAIC replaces the generic caption with a structured counterfactual prompt that explicitly opposes the identity attributes of the original subject while leaving all action and scene attributes intact.

Attribute Decomposition. The LMM extracts a structured attribute vector $\mathbf{a} = \mathbf{a}_{\text{id}} \cup \mathbf{a}_{\text{sem}}$ for each verified subject. The identity attribute set $\mathbf{a}_{\text{id}} = \{a_{\text{gender}}, a_{\text{age}}, a_{\text{build}}, a_{\text{tone}}, a_{\text{clothing}}\}$ captures biometrically salient properties that Re-ID systems exploit for identification. The semantic attribute set $\mathbf{a}_{\text{sem}} = \{a_{\text{action}}, a_{\text{scene}}, a_{\text{viewpoint}}\}$ captures properties required for downstream behavioral analysis and must be preserved unchanged.

Counterfactual Prompt Construction. To dissociate the synthetic persona from the source identity while ensuring demographic fairness, we construct a counterfactual prompt P_{cf} via a stochastic transformation. This process applies an augmented opposition function $\text{CF}(\cdot)$ to identity-bearing attributes $\mathbf{a}_{\text{id}}$, while preserving invariant semantic attributes $\mathbf{a}_{\text{sem}}$ (e.g., actions and scene). Formally:

$$P_{\text{cf}} = \text{Compose}\left(\{\text{CF}(a_i, \xi)\}_{a_i \in \mathbf{a}_{\text{id}}}, \{a_j\}_{a_j \in \mathbf{a}_{\text{sem}}}\right) \tag{5}$$

where ξ represents a random variable sampled from a categorical distribution to ensure diversity in skin tone, gender, and ethnicity. For discrete attributes, $\text{CF}(\cdot)$ maps to a categorical complement; for ordinal attributes, it maps to the opposing end of the spectrum. Critically, the injection of ξ-driven randomness prevents the model from converging toward a singular biased demographic, fostering a balanced representation across the generated output. For instance, a source description *"a young woman in a red jacket walking through a museum"* yields P_{cf} = *"an elderly man in a dark jacket walking through a museum"*, where every identity attribute changes while the action and scene are retained.

3.3 Stage III: Constrained Identity-Suppressive Generation

The generative backbone addresses F3 by synthesizing temporally coherent, identity-suppressed video conditioned jointly on the privacy mask from Stage I and the counterfactual prompt from Stage

II. We treat the task as a prompt-guided, mask-constrained video inpainting problem solved by a Latent Diffusion Transformer (DiT) Peebles & Xie (2023).

Generative Backbone. The input video is encoded into a latent representation z_0 by a pre-trained VAE Rombach et al. (2022). The spatiotemporal latent volume is divided into non-overlapping patches of size $p \times p$ and flattened into a token sequence X_{seq} following Video LDMs Blattmann et al. (2023), enabling the model to capture long-range temporal dependencies required for action continuity.

Tripartite Conditioning. The denoising process is conditioned on three complementary signals. *Structural conditioning* provides the masked video latent $z_{\text{ctx}} = \mathcal{E}(V \odot (1 - M))$ and the binary mask latent m_{lat}, concatenated with the noisy latent z_t. By occluding the identity region and providing the surrounding context, the model is constrained to synthesize content that is consistent with the scene lighting, background motion, and spatial layout rather than reconstructing the original subject. *Semantic conditioning* encodes the counterfactual prompt P_{cf} into text embeddings y_{text} via a frozen text encoder, injected through cross-attention layers. Unlike prior work that uses a generic caption, P_{cf} provides explicit counterfactual identity direction while preserving action semantics, guiding the model toward the specified counterfactual identity. *Temporal conditioning* is achieved implicitly through the DiT's full spatiotemporal attention over the token sequence X_{seq}, which enforces cross-frame coherence and produces a stable synthetic identity throughout the sequence, directly addressing F3.

Inference Procedure. Rather than training the model from scratch, we adopt a pretrained DiT-based diffusion model and perform inference directly. Given the masked region, the model synthesizes content conditioned jointly on the surrounding context z_{ctx} and the counterfactual prompt y_{text}, producing a plausible, identity-suppressed replacement that integrates naturally into the original video while conforming to the target counterfactual identity specified by Stage II. The denoising process follows the same formulation as the original training objective:

$$\hat{\epsilon} = \text{DiT}_\theta(z_t,\, t,\, \text{Concat}(z_{\text{ctx}}, m_{\text{lat}}),\, y_{\text{text}}) \tag{6}$$

with θ fixed to the pretrained checkpoint (no gradient updates applied).

4 Experiments

We evaluate MOSAIC using the privacy–utility evaluation protocol (Protocol(Sec. 4.2), assessing biometric suppression against pre-trained adversaries alongside semantic and perceptual utility.

4.1 Experimental Setup

Datasets. To ensure robust evaluation across divergent viewing paradigms, we conduct experiments on two complementary datasets:

OpenHumanVid (OHV) Li et al. (2025): OHV is a large-scale and high-quality human-centric video dataset. Serving as our standard benchmark for third-person perspectives, this large-scale dataset features diverse human activities in-the-wild, providing a rigorous testbed for preserving complex kinematic semantics.

Belvedere Museum (Ours): To validate real-world robustness in ego-centric scenarios, we introduce a novel dataset curated in collaboration with the Belvedere Museum, Vienna. Captured via wearable Tobii Pro Glasses Tobii AB, this benchmark introduces unique challenges: rapid motion blur from head movements, extreme dynamic range (spotlights vs. shadows), and the semantic necessity to strictly distinguish biological visitors from artistic depictions. Because the recordings contain naturalistic visitor behavior rather than staged participants, we treat whole-body anonymization as part of a broader responsible data-sharing pipeline rather than as a standalone legal guarantee.

The dataset comprises two continuous raw egocentric recordings with synchronized eye-tracking gaze data. To facilitate research into generative anonymization, we partitioned these recordings into 618 short video segments. From this corpus, we curated a representative test subset featuring diverse camera scales—including full-body, upper-body, and close-up views—and a broad demographic

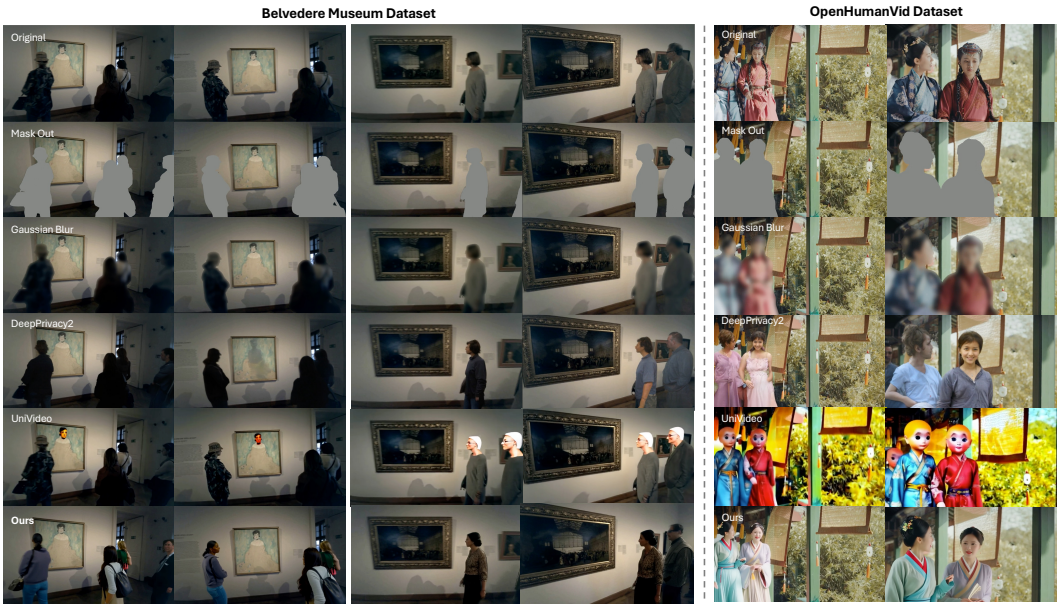

Figure 3: **Qualitative results on the Belvedere Museum and OpenHumanVid datasets.** We compare the visual performance of the Original, Mask Out, Gaussian Blur, DeepPrivacy2, and UniVideo methods against our proposed method (Ours).

range spanning various ages and genders. This subset specifically targets challenging edge cases, such as the presence of sculptures, paintings, and significant motion blur. To support comprehensive analysis, we provide both the raw segmentation masks and the masked video data, establishing a robust baseline for generative anonymization tasks.

Baselines and Implementation. We compare against two task-matched generative baselines: *DeepPrivacy2* Hukkelås & Lindseth (2023) , used in its full-body anonymization mode, and *UniVideo* Wei et al. (2026), a general-purpose video-to-video editing model prompted for whole-body identity replacement. We additionally report two non-generative references that expose the extremes of the privacy–utility trade-off: *Gaussian Blur* (perturbation) and *Mask Out* (removal). These references are not competing GVA methods, since neither produces a photorealistic, semantically equivalent subject. For privacy evaluation, we employ a frozen OSNet architecture Zhou et al. (2021); action consistency is measured with features from a pretrained R(2+1)D network Tran et al. (2018).

4.2 Evaluation Protocol

We evaluate the GVA framework across two competing dimensions: Identity Privacy and Video Utility, quantifying the trade-off between biometric suppression and semantic preservation.

To assess **Identity Privacy**, we simulate a re-identification (ReID) attack using a pre-trained OSNet Zhou et al. (2021). We utilize **Rank-1 accuracy** and **mAP** to determine if an anonymized subject can be successfully retrieved from a gallery of known identities. Furthermore, we monitor the **Cosine Similarity** ($\mathcal{S}_{cos}$) between source and anonymized embeddings to quantify the displacement in the identity latent space; lower values indicate that the biometric signatures have been effectively decoupled.

Video Utility is evaluated based on the video's downstream analytical value and perceptual realism. The **Scene Preservation Score (SPS)** is designed to ensure environmental integrity by measuring the structural stability of background regions unaffected by the subject transformation. To penalize common generative artifacts such as flickering, the **Temporal Warping Error (TWE)** assesses the consistency of motion dynamics via inter-frame optical flow discrepancy. Finally, **Action Consistency** ($\mathcal{A}_{cons}$) verifies the retention of behavioral semantics using R(2+1)D Tran et al. (2018) features, ensuring the anonymized content remains viable for downstream action recognition tasks.

Table 1: **Quantitative results on OpenHumanVid.** Gaussian Blur and Mask Out are non-generative references; highlighting marks the best value among generative methods only. The generative methods occupy different privacy–utility operating points; none is best on all metrics.

Method	Privacy Performance			Utility Performance		
	$\mathcal{S}_{cos}\downarrow$	R-1 (%)$\downarrow$	mAP (%)$\downarrow$	SPS$\uparrow$	TWE$\downarrow$	$\mathcal{A}_{cons}\uparrow$
Original	1.000	100.00	100.00	1.000	0.00	100.00
Blur	0.583	32.40	25.21	0.985	1.56	54.65
Mask Out	0.108	0.56	0.94	0.984	3.40	36.99
DeepPrivacy2 (body)	0.743	66.60	54.68	0.977	15.43	48.78
UniVideo	0.706	75.00	54.65	0.780	0.45	76.29
MOSAIC (Ours)	**0.660**	**59.96**	**42.34**	**0.878**	**17.54**	**73.32**

Table 2: **Evaluation on the Belvedere Museum dataset (Ego-centric).** Grouping and highlighting as in Table 1. MOSAIC attains the lowest identity leakage among generative methods on all three privacy metrics; utility rankings are metric-dependent.

Method	Privacy Performance			Utility Performance		
	$\mathcal{S}_{cos}\downarrow$	R-1 (%)$\downarrow$	mAP (%)$\downarrow$	SPS$\uparrow$	TWE$\downarrow$	$\mathcal{A}_{cons}\uparrow$
Original	1.000	100.00	100.00	1.000	0.00	100.00
Blur	0.723	61.58	32.69	0.990	219.73	92.52
Mask Out	0.455	20.23	10.79	0.990	890.48	82.95
DeepPrivacy2 (body)	0.715	67.13	26.95	0.980	397.89	88.72
UniVideo	0.749	84.17	42.24	0.808	95.47	80.77
MOSAIC(Ours)	**0.697**	**41.24**	**17.41**	**0.815**	**183.35**	**84.87**

4.3 Quantitative Analysis

We evaluate the performance of *MOSAIC* against several baselines on the OpenHumanVid (Table 1) and Belvedere Museum (Table 2) datasets, focusing on the trade-off between identity suppression and semantic preservation.

Identity Suppression. Among the generative methods, *MOSAIC* attains the lowest identity leakage on all three privacy metrics on both datasets. On OpenHumanVid, it obtains Rank-1/mAP of 59.96/42.34 versus 66.60/54.68 for *DeepPrivacy2* and 75.00/54.65 for *UniVideo*; on Belvedere, the corresponding values are 41.24/17.41 versus 67.13/26.95 and 84.17/42.24, a 25.89 pp Rank-1 reduction relative to *DeepPrivacy2*. The non-generative references mark the extremes of the trade-off. *Mask Out* removes the subject entirely and yields the lowest leakage; on OpenHumanVid, *Gaussian Blur* also yields lower Rank-1 than *MOSAIC* (32.40 vs. 59.96). However, neither produces a photorealistic, semantically equivalent subject, as reflected in their action consistency on OpenHumanVid (36.99% and 54.65%, vs. 73.32% for *MOSAIC*).

Semantic Preservation and Utility. The utility of the anonymized video is measured through scene fidelity (SPS), temporal consistency (TWE), and action consistency ($\mathcal{A}_{cons}$). *Scene fidelity:* while *UniVideo* serves as a strong generative baseline, its SPS scores (0.780 on OHV and 0.808 on Belvedere) are lower than ours (0.878 and 0.815), consistent with our qualitative observation that general-purpose models introduce stylized artifacts and fail to preserve background integrity. *Temporal stability:* in high-motion egocentric sequences, *DeepPrivacy2* suffers from high TWE (397.89 on Belvedere), reflecting its jittery synthesis, whereas *MOSAIC* more than halves this error (183.35); *UniVideo* attains the lowest TWE, partly because it often under-edits the subject. *Action consistency:* our method achieves competitive $\mathcal{A}_{cons}$ (73.32% on OHV and 84.87% on Belvedere), comparable to *UniVideo* (76.29% and 80.77%), while providing substantially stronger privacy protection.

In summary, across both datasets and all three privacy metrics, *MOSAIC* achieves the strongest identity suppression among the evaluated generative methods (on Belvedere, Rank-1 41.24 vs. 67.13 for DeepPrivacy2 and 84.17 for UniVideo). Its utility is competitive but metric-dependent: UniVideo attains lower TWE, and DeepPrivacy2 attains higher SPS on both datasets. We therefore characterize these methods as occupying different privacy–utility operating points rather than a single optimum.

Table 3: **Ablation study on the Belvedere Museum dataset.** We analyze the effectiveness of each agent component. Our full framework achieves the most balanced profile across privacy and utility.

Method	Privacy Performance			Utility Performance		
	$\mathcal{S}_{cos}\downarrow$	R-1 (%) $\downarrow$	mAP (%) $\downarrow$	SPS $\uparrow$	TWE $\downarrow$	$\mathcal{A}_{cons}\uparrow$
w/o Stage I	0.652	38.15	15.20	0.542	210.45	78.30
w/o Stage II	0.765	68.40	35.12	0.810	185.20	85.10
w/o Stage III	0.710	45.30	20.50	0.790	402.15	72.40
MOSAIC (Full)	**0.697**	**41.24**	**17.41**	**0.815**	**183.35**	**84.87**

4.4 Ablation Study

We conduct an ablation study on the Belvedere Museum dataset to evaluate the contribution of each MOSAIC stage. Removing *the Grounding Agent (Stage I)* leads to indiscriminate segmentation and over-anonymization, causing the SPS to drop sharply to 0.542 as the model fails to distinguish humans from static art. *The Planning Agent (Stage II)* is vital for privacy; its absence results in significant identity leakage, with Rank-1 accuracy rising to 68.40%. Finally, omitting the *Constrained Generation stage (Stage III)* collapses temporal stability, evidenced by a TWE spike to 402.15. Overall, the full MOSAIC framework provides the most balanced profile across privacy and utility metrics.

4.5 Qualitative Evaluation

Figure 3 illustrates a visual comparison between our proposed method and several state-of-the-art baselines. While traditional *Mask Out* methods adversely affect scene context and social utility, *Gaussian Blur* leaves structural features and color signatures visible. Among generative baselines, *DeepPrivacy2* produces high-fidelity individual frames but exhibits temporal inconsistencies, such as identity flickering and jittery attribute synthesis, largely due to its frame-independent processing and heavy reliance on precise human detection. Furthermore, while *UniVideo* represents a state-of-the-art general video generation model, it struggles with semantic misalignment in the anonymization task; despite extensive prompt engineering, it often under-edits the subject or generates stylized, cartoonish artifacts and fails to maintain background stability. In contrast, our method more consistently completes whole-body replacement while preserving the background. By leveraging optical flow-guided propagation and structure-aware conditioning, our framework synthesizes plausible, identity-replaced figures that maintain stable interaction geometry and gaze direction, yielding a coherent visual narrative.

5 Discussion and Broader Impact

Societal impact and responsible release. GVA is motivated by the need to share useful video data while reducing whole-body biometric leakage in any setting where recordings may contain identifiable people. This concern is broadly relevant across legal and institutional contexts, including but not limited to data protection frameworks such as the GDPR(General Data Protection Regulation) Voigt & Von dem Bussche (2017). We emphasize that generative anonymization is a technical safeguard rather than a legal guarantee of anonymity: residual re-identification risk depends on the threat model, auxiliary information, and downstream release setting. This is especially important for naturalistic datasets such as Belvedere Museum, where tourists and bystanders may appear without a controlled recording setup. Responsible release should therefore combine GVA with dataset documentation, access control for raw data, evaluation against adaptive Re-ID adversaries, and clear usage restrictions. Potential risks include deceptive video editing, overreliance on anonymization, and uneven protection across demographics, clothing, motion patterns, and lighting conditions.

Future direction. Future work will focus on enhancing robustness against adaptive Re-ID adversaries and investigating finer-grained decoupling between identity-bearing gait and action semantics. We also intend to address demographic fairness to ensure uniform protection across diverse populations and environmental conditions. Additionally, exploring long-term temporal anchoring mechanisms could further mitigate potential identity drift in extended video sequences.

AI USE STATEMENT

In this work, we used generative AI tools to search for and summarize relevant literature and to edit the manuscript for clarity and readability. We have not used generative AI tools to generate synthetic data, develop theoretical models, formulate or prove mathematical claims, propose hypotheses, design the method or experiments, implement the method, translate text, clean datasets, conduct qualitative analysis, or interpret experimental results. Suggested references were checked against the cited papers, and the edited prose was checked against our technical claims and experimental results. We take responsibility for the final content of this work, including any text produced with the aid of generative AI.

REFERENCES

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22563–22575. IEEE, 2023.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 8126–8133, 2019.

Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European conference on computer vision*, pp. 640–658. Springer, 2022.

Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pp. 363–370. Springer, 2003.

Michal Geyer, Omer Bar Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *International Conference on Learning Representations*, volume 2024, pp. 1608–1620, 2024.

Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1050–1059. IEEE, 2022.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 14993–15002. IEEE, 2021.

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

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

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15908–15918. IEEE, 2023.

Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. Dcface: Synthetic face generation with dual condition diffusion model. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12715–12725. IEEE, 2023.

Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, et al. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7752–7762. IEEE, 2025.

Xiang Li, Ancong Wu, and Wei-Shi Zheng. Adversarial open-world person re-identification. In *European Conference on Computer Vision*, pp. 287–303. Springer, 2018.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22511–22521. IEEE, 2023.

Jiang Liu, Chun Pong Lau, Zhongliang Guo, Yuxiang Guo, Zhaoyang Wang, and Rama Chellappa. Diffprotect: Generate adversarial examples with diffusion models for facial privacy protection. *arXiv preprint arXiv:2305.13625*, 2023.

Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. Ciagan: Conditional identity anonymization generative adversarial networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5446–5455. IEEE, 2020.

Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Faceless person recognition: Privacy implications in social media. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2016.

Stanley Osher and James A Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations. *Journal of computational physics*, 79(1):12–49, 1988.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Michael Ryoo, Brandon Rothrock, Charles Fleming, and Hyun Jong Yang. Privacy-preserving human activity recognition from extreme low resolution. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and effective obfuscation by head inpainting. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5050–5059. IEEE, 2018.

Tobii AB. Tobii pro lab. Computer software. URL `http://www.tobii.com/`.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 6450–6459. IEEE, 2018.

Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr): a practical guide. 2017.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Zhaofa Wang, Liyang Wang, Zhiping Shi, Miaomiao Zhang, Qichuan Geng, and Na Jiang. A survey on person and vehicle re-identification. *IET Computer Vision*, 18(8):1235–1268, 2024b.

Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. Univideo: Unified understanding, generation, and editing for videos. In *International Conference on Learning Representations*, volume 2026, pp. 113905–113933, 2026.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7589–7599. IEEE, 2023.

Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44 (6):2872–2893, 2021.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824. IEEE, 2023.

Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3701–3711. IEEE, 2019.

Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Learning generalisable omni-scale representations for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 5056–5069, 2021.

A Detailed Evaluation Protocol

In this section, we provide the formal mathematical definitions for the evaluation metrics used to assess the MOSAIC framework, as introduced in Section 4.2. Let $V = \{I_1, I_2, \ldots, I_T\}$ denote the original video sequence of length T, and $\hat{V} = \{\hat{I}_1, \hat{I}_2, \ldots, \hat{I}_T\}$ denote the corresponding anonymized sequence generated by MOSAIC.

A.1 Identity Privacy Metrics

To quantify the degree of biometric suppression, we utilize metrics grounded in the feature space of a pre-trained Re-identification (ReID) model. Let $\Phi_{\text{OSNet}}(\cdot)$ represent the feature extraction function of the OSNet Zhou et al. (2019) backbone. li2018adversarial

Rank-1 Accuracy and mAP We simulate a re-identification attack by treating the anonymized frames $\hat{I}_t$ as the query set and the original frames I_t as the gallery.

- **Rank-1 Accuracy** measures the probability that the top-1 retrieved match from the gallery belongs to the same identity as the query. In the context of MOSAIC, a lower Rank-1 accuracy indicates superior identity obfuscation.
- **Mean Average Precision (mAP)** provides a comprehensive measure of retrieval performance across all query samples. A reduction in mAP signifies that the unique biometric signatures have been effectively disrupted across multiple feature dimensions.

Cosine Similarity ($\mathcal{S}_{cos}$) To measure the displacement of identity embeddings in the latent space, we calculate the average cosine similarity between the source and anonymized features:

$$\mathcal{S}_{cos} = \frac{1}{T}\sum_{t=1}^{T} \frac{\Phi_{\text{OSNet}}(I_t) \cdot \Phi_{\text{OSNet}}(\hat{I}_t)}{\|\Phi_{\text{OSNet}}(I_t)\|_2 \|\Phi_{\text{OSNet}}(\hat{I}_t)\|_2} \tag{7}$$

A lower (or negative) value of $\mathcal{S}_{cos}$ indicates that the biometric features of the subject have been successfully decoupled from the original identity.

A.2 Video Utility Metrics

Video utility evaluates whether the anonymized video retains structural integrity, temporal smoothness, and semantic information for downstream tasks.

Scene Preservation Score (SPS) SPS measures the stability of background regions that should remain unaffected by the de-identification process. Let $M_t \in \{0,1\}^{H\times W}$ be a binary mask where pixels belonging to the subject are 1 and background pixels are 0. We define SPS as the Structural Similarity Index (SSIM) computed over the masked background:

$$\text{SPS} = \frac{1}{T}\sum_{t=1}^{T} \text{SSIM}\left(I_t \odot (1 - M_t), \hat{I}_t \odot (1 - M_t)\right) \tag{8}$$

where $\odot$ denotes the element-wise product. High SPS values indicate that the environmental context is preserved.

Temporal Warping Error (TWE) TWE is employed to penalize temporal artifacts such as flickering or inconsistent motion. Given the backward optical flow $O_{t\to t+1}$ from frame $t+1$ to t, the warping operation $\mathcal{W}(\hat{I}_{t+1}, O_{t\to t+1})$ estimates frame t from $t+1$. TWE is defined as the L_1 norm of the warping residual:

$$\text{TWE} = \frac{1}{T-1}\sum_{t=1}^{T-1} \frac{1}{N} \left\| \hat{I}_t - \mathcal{W}(\hat{I}_{t+1}, O_{t\to t+1}) \right\|_1 \tag{9}$$

where N is the total number of pixels. A lower TWE reflects higher temporal consistency and visual smoothness.

Action Consistency ($\mathcal{A}_{cons}$) To ensure the utility of the anonymized video for action recognition, we measure the preservation of behavioral semantics. Let $\Psi_{\text{R(2+1)D}}(\cdot)$ denote the spatiotemporal feature extractor from a pre-trained R(2+1)D Tran et al. (2018) network. Action consistency is defined as:

$$\mathcal{A}_{cons} = \frac{\Psi_{\text{R(2+1)D}}(V) \cdot \Psi_{\text{R(2+1)D}}(\hat{V})}{\|\Psi_{\text{R(2+1)D}}(V)\|_2 \|\Psi_{\text{R(2+1)D}}(\hat{V})\|_2} \tag{10}$$

A high $\mathcal{A}_{cons}$ score demonstrates that the MOSAIC framework maintains the subject's kinematics and behavioral intent.

B Related Work

B.1 Generative Identity Obfuscation

The paradigm of visual privacy has shifted from rudimentary obfuscation (e.g., blurring) to generative synthesis, aiming to reconcile privacy protection with data utility. Early adversarial frameworks Sun et al. (2018) and GAN-based architectures, such as DeepPrivacy Hukkelås et al. (2019) and CIAGAN Maximov et al. (2020), pioneered the synthesis of identity-agnostic faces while retaining pose and gaze attributes. Recently, diffusion models have emerged as the state-of-the-art, with methods like DC-Face Kim et al. (2023) and DiffProtect Liu et al. (2023) leveraging the semantic latent space of LDMs for high-fidelity controllable inpainting.

However, these approaches predominantly operate at the image level and focus narrowly on facial regions. They neglect the *whole-body biometric signature*—including gait, clothing layout, and body shape—which contemporary Person Re-Identification (Re-ID) systems effectively exploit to track individuals across non-overlapping views Ye et al. (2021); Wang et al. (2024b). Consequently, face-centric de-identification remains insufficient against modern surveillance analytics that utilize global biometric invariants.

B.2 Video Privacy and Temporal Consistency

Extending anonymization to the video domain introduces the challenge of temporal consistency, where frame-independent processing results in severe flickering and semantic drift. While DeepPrivacy2 Hukkelås & Lindseth (2023) attempts to mitigate facial instability via tracking-guided generation, comprehensive whole-body protection in video remains under-explored. Existing solutions primarily rely on adversarial perturbations Li et al. (2018), which aim to mislead classifiers imperceptibly but fail to remove perceptible identity cues required for human inspection.

Concurrent advancements in video-to-video editing Wu et al. (2023); Geyer et al. (2024); **?**); Khachatryan et al. (2023) have demonstrated impressive attribute modification capabilities. However, these general-purpose frameworks lack the precise *privacy-utility disentanglement* required for surveillance tasks. They frequently induce motion hallucinations or alter background semantics (e.g., changing the weather or scene layout), rendering the data unusable for behavioral analysis. In contrast, our ISVT framework targets the holistic suppression of identity features while rigorously grounding the generative process in the original action dynamics and environmental context.

B.3 Visual Grounding and Structural Control

Our approach is further inspired by recent advances in spatially grounded generation and video object segmentation. In static domains, ControlNet Zhang et al. (2023) and GLIGEN Li et al. (2023) have introduced robust mechanisms to condition diffusion models on structural primitives (e.g., edges, depth). Adapting these constraints to video necessitates precise temporal correspondence, a task partially addressed by tracking-based segmentation models Cheng & Schwing (2022); Yang et al. (2023). Nevertheless, these tools are not inherently privacy-aware; they lack the semantic reasoning to distinguish between sensitive biological subjects (e.g., a visitor) and non-sensitive artistic depictions (e.g., a statue). We bridge this gap by integrating LMM-guided semantic reasoning with structure-aware generation, ensuring robust and precise anonymization.

C Computational Resources and Efficiency

C.1 Hardware and Runtime Configuration

All inference and evaluation experiments were conducted on a single NVIDIA A100 GPU (80GB VRAM variant). Our implementation utilizes the PyTorch framework with FP16 precision to balance computational throughput and generative fidelity.

C.2 Inference Efficiency

For a standard input sequence of 5 seconds at 30 FPS (totaling 150 frames), the peak VRAM usage for the full MOSAIC pipeline is approximately **24 GB**. This memory footprint covers the following concurrent processes:

- **Grounding Agent:** Multi-object tracking and mask propagation via dense optical flow and SDF interpolation.
- **Semantic Reasoning:** The shared LMM call for track-based filtering and counterfactual identity planning.
- **Generative Backbone:** Denoising via the Latent Diffusion Transformer (DiT) with tripartite conditioning.

C.3 LMM Call Reduction

As established in Section 3.1, a significant challenge in agentic video generation is the high latency of Large Multimodal Models (LMMs). By adopting a track-conditioned strategy where the LMM is invoked only upon the first observation of a unique track ID, MOSAIC reduces the number of LMM calls from $\mathcal{O}(T)$ (per-frame) to $\mathcal{O}(N_{tracks})$ (per-track). In our benchmarks on the Belvedere Museum dataset, this results in a reduction of computational overhead for the reasoning stage by over 90% on average.