ReGenHuman: Re-Generating Human Appearances for Realistic Full-Body Video Anonymization

Adam Sun[*], **Eshaan Barkataki, Arnold Milstein, Gordon Wetzstein, Ehsan Adeli**
Stanford University

Abstract

Anonymizing human-centric video data is an understudied problem. Prior anonymization techniques either blur or redact pixels at the cost of realism and downstream utility, or generate frame-by-frame at the cost of temporal coherence. We introduce **ReGenHuman**, the first full-body video anonymization pipeline that is simultaneously realistic, temporally consistent, and anonymous by construction. Contrary to past approaches which redact or edit the inputs directly, we propose a *regenerate, don't edit* paradigm. Our approach composites 2D pose, segmentation, and monocular depth into two complementary conditioning streams—**StructAll** and **StructHuman**, which are used to fine-tune a video-to-video diffusion backbone on in-the-wild human videos, synthesizing the human regions entirely from identity-free structural cues. We evaluate our model on privacy, quality, and utility, and show that our **ReGenHuman** achieves the best tradeoff across all three axes against current baselines. We further show that our anonymized videos remain effective for downstream tasks, including video question answering. The project page can be found at `regenhuman.github.io`.

1 Introduction

Modern vision systems rely on large, human-centric video datasets to power applications ranging from action recognition and video understanding [7, 13, 28] to autonomous vehicles [15, 59–61] and video surveillance [58, 63, 45, 44]. Data is especially important in clinical settings, where bedside monitoring and ambient intelligence applications require diverse, multi-person footage captured in real-world environments [37, 65, 11, 8, 1, 49]. However, raw videos present a severe privacy risk; identity cues are apparent not only in faces and bodies but also in clothing, gait, and background surroundings [60, 66, 50]. Consequently, regulations such as the US Health Insurance Portability and Accountability Act (HIPAA) [2] and the General Data Protection Regulation [46] require consent that is effectively impossible to obtain at the scale. This challenge is compounded by the fact that trained models themselves are vulnerable to privacy breaches. Attacks such as membership inference can recover identifying frames from a model's weights alone [43, 33, 21].

The sensitive domains that stand to benefit most from large-scale video learning—such as hospitals, classrooms, and workplaces—are precisely the areas where data cannot be openly released. As a result, AI performance in these critical settings falls behind [19, 42]. The natural solution is anonymization: transforming a video so that no individual can be re-identified, while preserving the semantic content that downstream models depend on. Text [18, 36, 39] and image [26, 24, 25, 14] anonymization have been extensively studied; however, video anonymization remains underexplored.

Following past work [14, 60], we argue that a good video anonymizer must satisfy three objectives *simultaneously*: **(P1) Privacy.** The output must not be re-identifiable by any means, including through face, body, clothing, and background cues. **(P2) Quality.** The video must be temporally and

[*]Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

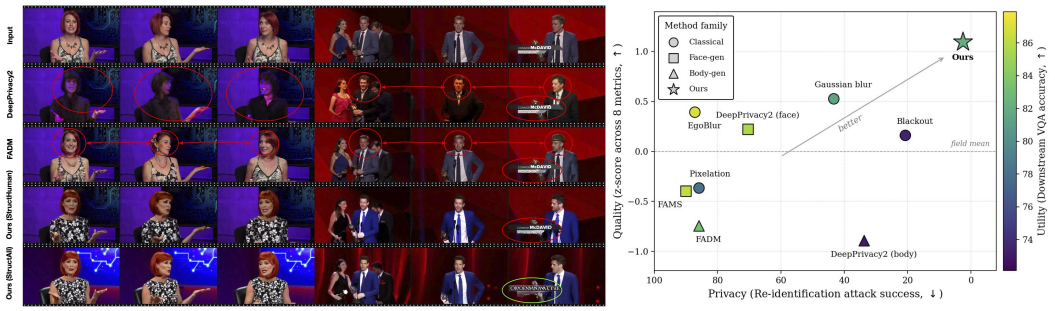

Figure 1: We propose **ReGenHuman**, the first pipeline for realistic full-body video anonymization. Past methods for full-body anonymization such as DeepPrivacy2 [24] and FADM [66] are inadequate for video — they either generate anatomically unrealistic humans or do not depict the same identity across frames. On the other hand, our **ReGenHuman**'s generations are realistic and temporally consistent, with our **StructAll** generation being able to anonymize the background and accompanying text (*left*). We evaluate current anonymization approaches on privacy (Clothes-changing re-identification mAP on CCVID [16] with CSCI [38]), quality (z-score across metrics from VBench 1 and 2 [23, 62]), and utility (zero-shot video question answering accuracy on HOI-Gen1M [34]). Our approach achieves the best tradeoff of all prior baselines, setting a new standard in video anonymization (*right*).

geometrically consistent, and realistic enough that models trained on real-world videos can utilize it without distribution shift. **(P3) Utility.** Actions, poses, human–object interactions, and scene context must remain recognizable, preserving the semantic structure required by downstream tasks.

The past paradigm for anonymization mainly involves *redacting, editing or perturbing* the original pixels—a paradigm that either removes useful information or does not guarantee full anonymization coverage. We propose the opposite paradigm: *regenerate* the human region from scratch, conditioned only on identity-free structural signals.

Our ReGenHuman pipeline consists of three distinct steps. Firstly, we extract pixel-precise human masks [9], 2D pose skeletons [56], depth [10], and a detailed text caption from the input video [34, 4]. Secondly, we combine the visual conditionings into two structural conditioning videos **StructAll** and **StructHuman**, which are fully anonymous and preserve the context and dynamics of the original video. Lastly, we fine-tune a pretrained video-to-video diffusion model [47, 27] on real-world human videos from the HOIGen-1M dataset [34]. At inference, our trained anonymizer produces realistic, context-consistent anonymizations of input videos that are still useful for downstream applications, outperforming past anonymization approaches in privacy, quality, and utility.

Our contributions are the following:

- We present the first realistic and temporally consistent video anonymization pipeline. By jointly denoising full clips conditioned on structural overlays, our method produces anonymized video that is visually realistic and contextually consistent with the input video.

- We propose a new anonymization paradigm: regenerate, don't edit. Unlike all prior methods, which edit, redact, or perturb the original pixels, we regenerate the human region from identity-free structural inputs — our **StructAll** and **StructHuman**. This paradigm shift carries an architectural privacy guarantee: no original human pixel can reach the output.

- We outline a unified three-axis evaluation protocol for full-body video anonymization, jointly measuring **Privacy** (identity and text similarity, clothes-changing person re-identification), **Quality** (VBench- and VBench-2.0–aligned consistency, realism, and smoothness metrics), and **Utility** (video question answering).

We conduct detailed experiments showing that our method achieves the best privacy–quality–utility tradeoff among current anonymization approaches, including in the clinical domain. We hope this work inspires further research in the field of video anonymization.

Table 1: Comparing current video anonymization methods. **Fully Private**: full-body coverage *and* an architectural guarantee that no original human pixel reaches the output. **Temporal**: output is temporally consistent (no per-frame flicker of identity). **Utility**: action, pose, HOI, and scene context are preserved. **Realistic**: output is a human-viewable video that is photorealistic and in-distribution for models trained on regular video. Ours is the only method that satisfies all four.

Method	Fully Private	Temporal	Utility	Realistic
EgoBlur [40]	×	✓	✓	×
DeepPrivacy [26], FAMS [31]	×	×	✓	×
DeepPrivacy2 [24]	✓	×	✓	×
FADM [66], Unsafe2Safe [14]	×	×	✓	✓
STPrivacy [32], Pixels2Privacy [3]	×	✓	✓	×
ReGenHuman (ours)	✓	✓	✓	✓

2 Related Work

2.1 Image anonymization

The classical approach for image anonymization detects faces or bodies and blurs, pixelates, or masks out the regions. These pipelines remain the de-facto standard in deployed systems [40] and are simple and verifiable. However, once faces, skin, and clothing are redacted, action recognition, video question answering, and other downstream tasks degrade sharply [25, 60].

To improve the utility of anonymized images, generative anonymization instead replaces the redacted region with a synthesized one. For faces, GAN-based methods [26, 24, 41, 30] inpaint a new face from a bounding box or mask, and diffusion-based approaches such as Face Anonymization Made Simple (FAMS) [31] generate a plausible by face by partially noising and denoising the face region. These approaches may produce photorealistic faces but are insufficient for privacy: clothing, tattoos, and background remain intact, and modern body re-identification [16, 38, 64] re-links subjects from these cues even when the face is perfectly hidden [60].

A smaller line of work extends generative anonymization to the full body. DeepPrivacy2 [24] inpaints the human region from a mask, but its GAN-based generations frequently look unrealistic. FADM [66] uses a text-to-image diffusion model, but conditions the diffusion process on the original image and so offers no anonymization guarantee. Unsafe2Safe [14] proposes a multi-stage pipeline combining LLMs, VLMs, and diffusion image editors. However, the editor stage suffers from the same issue, receiving the original image and relying on an edit instruction.

Furthermore, the image anonymizers above do not transfer well to video. With no access to other frames, the synthesized identity drifts over time, evident in Figure 1 (*left*).

2.2 Video anonymization

Video-native methods are sparse in the literature. Tracked-redaction approaches such as FAKER [5] and the medical-domain MedVidDeID [37] blur or replace the person consistently across frames, avoiding flicker but inheriting the utility loss of pixel-level redaction. Privacy-preserving action recognition [32, 3] sidesteps pixel-level anonymization entirely by removing identity-bearing tokens from the input, but the output is not a human-viewable video and cannot be released for general downstream use. To the best of our knowledge, no generative video anonymization method exists.

2.3 Controllable video generation

Text-to-video diffusion models [6, 17, 57, 47] denoise an entire clip jointly, producing temporally coherent video from a text prompt. Subsequent work added structural conditioning—depth, 2D pose, segmentation masks, or Canny edges as per-frame control channels [55, 35, 53, 51, 47]—and unified video-to-video models such as Wan Fun Control [47] and VACE [27] accept an arbitrary mix of structural controls, masks, plus an optional reference image.

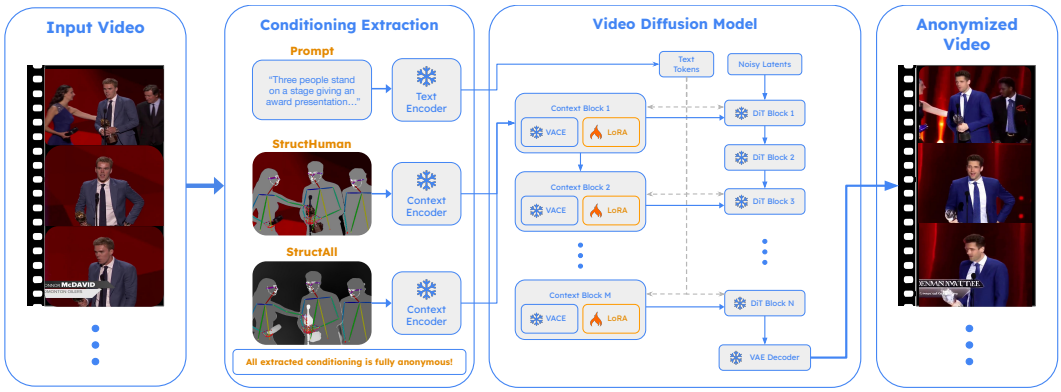

Figure 2: **ReGenHuman pipeline.** Given an input video, we extract fully anonymous conditioning: a text caption plus a StructHuman or StructAll conditioning video. We freeze both the DiT Blocks and the Context Blocks of VACE [27] and fine-tune a LoRA [20] inside each Context Block. At inference, the model produces the anonymized video *without ever seeing the input video*.

No prior anonymization method can anonymize a video consistently and realistically across frames. Furthermore, controllable video diffusion models were developed for creative generation and, to our knowledge, have never been applied for anonymization. Yet their two core properties—*joint temporal denoising* and *appearance-free structural conditioning*—are exactly the features that every prior video anonymizer was missing. We combine these insights and propose utilizing video-to-video diffusion models as full-body anonymizers.

3 Method

3.1 Wan VACE

Wan [47] is an open-source video diffusion model that has been pretrained on a large corpus of video data, making it generalizable for our use case. Let $\mathbf{x} = (x^1, \ldots, x^T)$ denote a video clip with 3D-VAE latent $\mathbf{z}_0 = \mathcal{E}(\mathbf{x})$ [47]. VACE [27] extends Wan for unified controllable video generation.

VACE works through two constructs. First, a *Video Condition Unit* (VCU) packs task conditioning into a single video-shaped tensor $\mathbf{c} = (c^1, \ldots, c^T) \in \mathbb{R}^{T \times H \times W \times 3}$ that may carry an arbitrary mix of structural signals (depth, pose, mask). Second, a stack of *Context Blocks* runs in parallel to the frozen DiT backbone, encoding the VCU with the same 3D VAE and injecting hidden states into every DiT layer. The rectified-flow denoiser becomes $v_\theta(\mathbf{z}_t, t, \mathcal{E}(\mathbf{c}), y)$ with text prompt y. The key property for anonymization is that the VCU is injected per-pixel, so we can hand the model a heterogeneous overlay without modifying the architecture.

3.2 Conditioning signal extraction

From an input clip $\mathbf{x}$, we extract three per-frame signals as full-length videos plus a clip-level caption y. Using state-of-the-art off-the-shelf models, we obtain a human-instance mask video $\mathbf{M} \in \{0,1\}^{T \times H \times W}$ from SAM3 [9], which propagates instance identities across frames; a rendered DWPose [56] skeleton video $\mathbf{P}$; and a monocular depth video $\mathbf{D}$ from Video Depth Anything [10], whose temporal-attention layers enforce frame-to-frame depth consistency.

Each of these three signals plays a distinct role and is appearance-free by construction. The human mask $\mathbf{M}$ tells the generator *where* a person must appear, outlining the exact region to generate. The depth map $\mathbf{D}$ anchors the person in the 3D scene and provides the necessary geometry and context of the scene without revealing identifying information such as text. Pose $\mathbf{P}$ ensures that the regenerated person is positioned in the same way as the source. Furthermore, the descriptive caption y ensures that the overall context of the clip remains consistent with the original.

We compose our signals into two conditioning videos. With Π^t denoting the pixels occupied by the poses, we define:

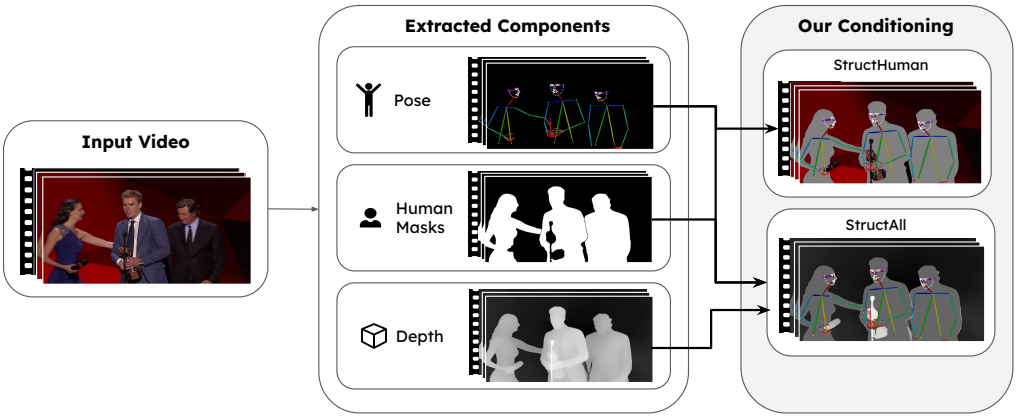

Figure 3: **Conditioning extraction.** We propose two types of conditioning, StructHuman and StructAll. Both replace the human with a flat fill and overlay the pose on top. StructHuman keeps the original background and objects, while StructAll replaces the background with the extracted depth.

StructHuman, which keeps non-human pixels (such as held objects and the background) non-anonymized when they carry no private information (e.g., public streets, generic hospital rooms). This approach trades scene anonymization for richer context.

$$c^t_{\text{Hum}} \;=\; \Pi^t \odot P^t \;+\; (1-\Pi^t) \odot \Big[\, M^t \odot \mathbf{g} \;+\; (1-M^t) \odot x^t \,\Big]. \tag{1}$$

StructAll, which uses depth as the background, yielding a fully-anonymized frame (no original pixels anywhere) while preserving scene geometry and the subject's pose:

$$c^t_{\text{All}} \;=\; \Pi^t \odot P^t \;+\; (1-\Pi^t) \odot \Big[\, M^t \odot \mu(D^t) \;+\; (1-M^t) \odot D^t \,\Big], \tag{2}$$

where $\mu(D^t)$ denotes a flat fill at the mean depth value of frame t, so the masked region sits at the average scene depth rather than introducing an arbitrary gray level into the depth map.

In both cases the human region is replaced by a flat, appearance-free fill before the pose is overlaid on top — neutral gray $\mathbf{g} = (128, 128, 128)$ for StructHuman and the mean-depth value $\mu(D^t)$ for StructAll. This prevents identifiable human features (including body shape encoded in the depth map) from leaking through the conditioning. Notice:

$$\forall t, \quad \big[c^t \odot M^t\big] \text{ depends on } P^t \text{ only, not on } x^t, \tag{3}$$

holds by construction, providing an architectural privacy guarantee: the diffusion model has no path through which the original human appearance can flow to the output.

We keep the entire video-to-video diffusion model—both the Wan DiT backbone and the pretrained VACE Context Block weights—*frozen*, and add rank-ρ=32 LoRA [20] factors A, B on the projection matrices of every Context Adapter block. We build on the lightweight 1.3B Wan2.1-VACE so that anonymizing a full dataset remains practical on a single GPU. This allows us to utilize the high-quality conditional generation capabilities of the pretrained model without dealing with catastrophic forgetting [29]. Writing $\mathbf{z}_1 = \mathcal{E}(\mathbf{x})$ for the VAE latent of the ground-truth clip, $\mathbf{z}_0 \sim \mathcal{N}(0, I)$, and $\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\mathbf{z}_1$ for the rectified-flow interpolant at $t \sim \mathcal{U}(0,1)$, the only trainable parameters $\theta = \{A, B\}$ are optimized with the conditional rectified-flow (flow-matching) objective

$$\mathcal{L}(\theta) = \mathbb{E}_{(\mathbf{x},\mathbf{c},y),\, t,\, \mathbf{z}_0}\big[\|v_\theta(\mathbf{z}_t, t, \mathcal{E}(\mathbf{c}), \tau(y)) - (\mathbf{z}_1 - \mathbf{z}_0)\|_2^2\big], \tag{4}$$

where v_θ is the velocity predicted by the LoRA-adapted network, $\mathbf{c} \in \{\mathbf{c}_{\text{All}}, \mathbf{c}_{\text{Hum}}\}$ is the structural conditioning video (encoded by the same VAE and consumed by the Context Adapter), and $\tau(y)$ is the umT5 embedding of the caption. StructHuman and StructAll are trained as two separate adapters, each on its own conditioning.

4 Experiments

4.1 Setup

For training our model, we utilize a randomly sampled 7,500-clip subset of HOIGen-1M [34], a corpus of over one million in-the-wild human–object interaction (HOI) videos curated from diverse sources for text-to-video HOI generation. The 1,000-clip evaluation subset used throughout Section 4.3 is sampled separately and is disjoint from the training clips. The data exemplifies the setting our anonymizer must handle in deployment: natural lighting, free-form camera motion, varied clothing, and multi-person, multi-object activities. The captions provided in the dataset are derived via a Mixture-of-Multimodal-Experts pipeline [34, 48, 54], maximizing descriptiveness and accuracy.

We implement classical baselines by performing pixel-level redactions applied inside the human masks $\mathbf{M}$, ranging from performing a Gaussian blur or pixelating to completely blacking out the human regions. We also follow the recommended implementations of EgoBlur [40], DeepPrivacy2 [24], FAMS [31], and FADM [66].

4.2 Evaluation protocol

Following past work [14, 60], we evaluate along three axes: privacy, quality, and utility, adapted to video. Each metric takes the triplet $(x, \hat{x}, y)$ (original clip, anonymized clip, caption) and returns a clip-level score.

Privacy. Two metrics measure how much of the original's identifying content survives in the anonymized clip. **Identity Similarity** [14] is the ArcFace cosine similarity between each detected face and its anonymized counterpart. **Text Similarity** is the mean best Levenshtein-ratio match between text detected in the original and in the anonymized clip. Neither measures how well the body is anonymized, so we also attack each method with a body-aware re-identifier: CSCI-V [38], a state-of-the-art video re-identification model, on the CCVID clothes-changing re-ID benchmark [16]. We anonymize a fixed 500-tracklet subset of the CCVID query split, keep the full gallery un-anonymized, and re-query; a match means the attacker recovered the identity despite anonymization.

Quality. For quality evaluation, we employ the widely used video generation evaluation benchmarks VBench [23] and VBench-2.0 [62]. From VBench, we report **Subject Consistency** and **Background Consistency** (within-clip coherence of the foreground subject and surrounding scene), **Motion Smoothness** and **Temporal Flickering** (frame-to-frame stability of motion and pixel intensity), and **Overall Consistency** (caption alignment). From VBench-2.0, we report **Human Anatomy** and **Human Identity**, which respectively measure anatomical plausibility and how stably the generated person persists across frames. We additionally prompt a VLM [12] on how likely the video is to be a real-world video and report a **Realism** score.

Utility. To test whether the anonymized clip still supports downstream tasks, we evaluate Video Question Answering in Section 4.4.

4.3 Results on HOIGen-1M

We anonymize a randomly sampled $1{,}000$-clip subset of HOIGen-1M [34] and report a qualitative comparison in Figure 6, as well as a privacy and quality comparisons in Table 2 and Table 3.

Face-only baselines leave the body untouched, resulting in high re-identification scores. Full body redactions that destroy the signal entirely (Blackout, pixelation, Gaussian blur) are much more private, but suffer in quality — they have low Realism scores and low Overall Consistency. Generative inpainting of the body suffers in quality and is still not completely private; FADM is even less private than the face-only DeepPrivacy2. Interestingly, classical baselines score high in some quality measures. We attribute this to the same redaction operation being performed across frames, trivially achieving high "consistency".

Our **StructAll** and **StructHuman** are the most private methods in the benchmark, driving re-identification to near chance under the identical matched protocol (StructAll 1.4 and StructHuman 3.0 clothes-changing R-1, against 92.0 with no anonymization). The fact that our method outperforms

Method	Appearance leakage		Body re-ID: Standard (SC)			Body re-ID: Clothes-Changing (CC)		
	Identity Similarity ↓	Text Similarity ↓	R-1 ↓	R-5 ↓	mAP ↓	R-1 ↓	R-5 ↓	mAP ↓
Classical								
EgoBlur (face only) [40]	0.0014	0.9263	100.0	100.0	99.9	86.6	91.2	87.2
Gaussian blur	0.0148	0.7694	50.0	60.8	53.3	40.0	51.6	43.3
Pixelation	0.0039	0.7549	100.0	100.0	99.8	84.8	90.6	85.9
Blackout	0.0039	0.7349	7.6	18.4	14.2	17.4	29.6	20.7
Face-only, generative								
FAMS [31]	0.0884	0.8396	100.0	100.0	100.0	90.8	92.0	90.1
DeepPrivacy2 (face) [24]	0.0603	0.9166	79.1	87.3	79.0	69.4	79.2	70.4
Body-only, generative								
DeepPrivacy2 (body) [24]	0.0309	0.7404	7.0	13.3	11.3	29.0	40.0	33.7
FADM [66]	0.0728	0.7961	96.2	98.7	96.7	86.4	90.4	85.9
Video-native overlays								
FAKER [5]	0.0014	0.7154	4.4	9.5	7.9	9.4	16.2	12.1
MedVidDeID [37]	0.0438	0.6053	84.8	96.2	85.9	56.8	69.2	58.6
Ours (Wan2.1-VACE conditioning)								
StructHuman (ours)	0.0092	0.3593	3.2	4.4	5.2	3.0	5.0	4.4
StructAll (ours)	0.0063	0.0753	3.2	3.2	4.6	1.4	2.0	2.5

Table 2: **Privacy evaluation.** Appearance-leakage metrics on HOIGen-1M [34] and body re-identification on CCVID [16]. Lower is better for all columns (↓). Per-column best and second highlighted.

even blackout on re-identification shows that re-generating the entire body with a realistic appearance not only removes identity signal but even misleads re-identification attempts. We lead on Overall Consistency, Human Identity, and Realism, are second only to FAKER on Human Anatomy, in addition to performing the best amongst the generative full body methods in the other quality metrics. Our method is also the only one to consistently anonymize text, a useful capability for many high-sensitivity scenarios.

Classical anonymizers (blackout, pixelation) may be private but collapse on quality. At the same time, anonymizers that may achieve a higher quality (EgoBlur, Gaussian blur) leave the body re-identifiable. Our methods sit near the top on both axes simultaneously, generating videos that are both private yet high quality.

4.4 Utility: zero-shot Video-QA and clinical fine-tuning

Anonymization quality only matters if the anonymized clip still answers the question the original clip would have. We evaluate Video Question Answering on three benchmarks: HOIGen-1M [34]; NExT-QA [52], a general-purpose multiple-choice video-QA benchmark; and MedVideoCap-55K [49], a caption-annotated medical-video corpus spanning surgery, bedside care, and patient-doctor interactions. We use a vision-language model (VLM) [4] to generate multiple choice questions from the provided detailed captions for HOIGen-1M and MedVideoCap-55K. For evaluation, a frozen VLM is prompted with a question in addition to 8 anonymized frames, and answers are scored by exact-match accuracy. Results are in Table 4.

Question answering results. Face-only baselines (EgoBlur, FAMS, DeepPrivacy2-face) sit within 0.9 points of the un-anonymized upper bound on all three benchmarks, with EgoBlur even slightly exceeding it on HOIGen ($+0.50$) and MedVideoCap ($+0.13$). This is unsurprising — they edit only the face, leaving the body pixels the VQA model relies on intact. Among methods that actually modify the body, only FADM beats us zero-shot, and the privacy table shows why: FADM's body re-ID is very high, suggesting that it is minimally modifying the video (see Figure 6). Our slight degradation in video question answering accuracy, on the other hand, is exchanged for privacy.

Performance on clinical data. Despite being trained only on the general HOIGen-1M [34] corpus, our model generalizes to clinical footage out of the box, as shown in Figure 4 and Table 4.

Method	Quality							
	Subject Consistency ↑	Background Consistency ↑	Motion Smoothness ↑	Temporal Flickering ↑	Overall Consistency ↑	Human Anatomy ↑	Human Identity ↑	Realism ↑
Classical								
EgoBlur (face only) [40]	92.18%	92.54%	96.92%	95.10%	12.43%	83.70%	13.20%	84.00%
Gaussian blur	92.02%	92.96%	97.41%	95.80%	13.62%	75.30%	14.49%	75.04%
Pixelation	93.01%	93.32%	96.44%	94.98%	13.14%	70.15%	6.93%	60.92%
Blackout	92.07%	92.42%	97.43%	95.96%	11.79%	81.13%	7.69%	69.76%
Face-only, generative								
FAMS [31]	90.71%	90.18%	96.61%	94.75%	12.16%	74.31%	8.23%	87.86%
DeepPrivacy2 (face) [24]	91.67%	91.26%	96.77%	94.88%	11.99%	87.34%	11.86%	88.45%
Body-only, generative								
DeepPrivacy2 (body) [24]	89.16%	88.76%	95.78%	94.03%	12.06%	76.17%	10.72%	83.74%
FADM [66]	88.19%	88.48%	95.86%	94.11%	12.02%	83.55%	10.88%	88.00%
Video-native overlays								
FAKER [5]	91.60%	91.74%	96.53%	94.68%	13.49%	89.83%	9.11%	46.42%
MedVidDeID [37]	90.68%	93.18%	96.95%	95.33%	13.31%	80.31%	18.49%	81.86%
Ours (Wan2.1-VACE conditioning)								
StructHuman (ours)	92.23%	91.52%	97.12%	95.05%	16.36%	89.74%	19.30%	89.49%
StructAll (ours)	92.88%	92.35%	96.98%	94.91%	16.11%	86.85%	17.35%	89.65%

Table 3: **Quality evaluation.** Quality [23, 62] metrics on HOIGen-1M [34]. Higher is better for all columns (↑). Per-column best and second highlighted.

Method	HOIGen (zero-shot)		NExT-QA (zero-shot)		MedVideoCap (zero-shot)	
	Acc. (%) ↑	Δ	Acc. (%) ↑	Δ	Acc. (%) ↑	Δ
Original						
Original (upper bound)	86.40	—	78.31	—	90.47	—
Classical						
EgoBlur [40]	86.90	+0.50	77.56	-0.75	90.60	+0.13
Gaussian blur	81.33	-5.07	72.93	-5.38	85.27	-5.20
Pixelation	78.20	-8.20	73.76	-4.55	85.13	-5.34
Blackout	73.47	-12.93	73.00	-5.31	74.80	-15.67
Face-only, generative						
FAMS [31]	86.13	-0.27	77.97	-0.34	89.60	-0.87
DeepPrivacy2 (face) [24]	85.60	-0.80	77.80	-0.51	89.80	-0.67
Body-only, generative						
DeepPrivacy2 (body) [24]	73.10	-13.30	73.39	-4.92	78.07	-12.40
FADM [66]	83.33	-3.07	76.49	-1.82	86.30	-4.17
Video-native overlays						
FAKER [5]	80.07	-6.33	69.32	-8.99	82.73	-7.74
MedVidDeID [37]	86.50	+0.10	72.98	-5.33	88.53	-1.94
Ours (Wan2.1-VACE conditioning)						
StructHuman (ours)	82.37	-4.03	75.37	-2.94	85.20	-5.27
StructAll (ours)	81.80	-4.60	75.85	-2.46	84.67	-5.80

Table 4: **Utility evaluation.** Zero-shot Video Question Answering accuracy with a frozen Qwen3-VL-8B [4] on anonymized clips. Δ is relative to the un-anonymized upper bound (top row); the upper bound is excluded from the per-column ranking. Per-column best and second highlighted.

This out-of-domain transfer matters because medical video carries a stricter privacy obligation than the in-the-wild footage we trained on. Our results suggest that the structural conditioning and caption carry enough scene context for the model to handle the medical setting without re-training.

4.5 Ablation Study

We ablate the components of our conditioning in Figure 5. Conditioning on depth alone produces geometrically correct but pose-incorrect generations, while conditioning on pose alone loses scene

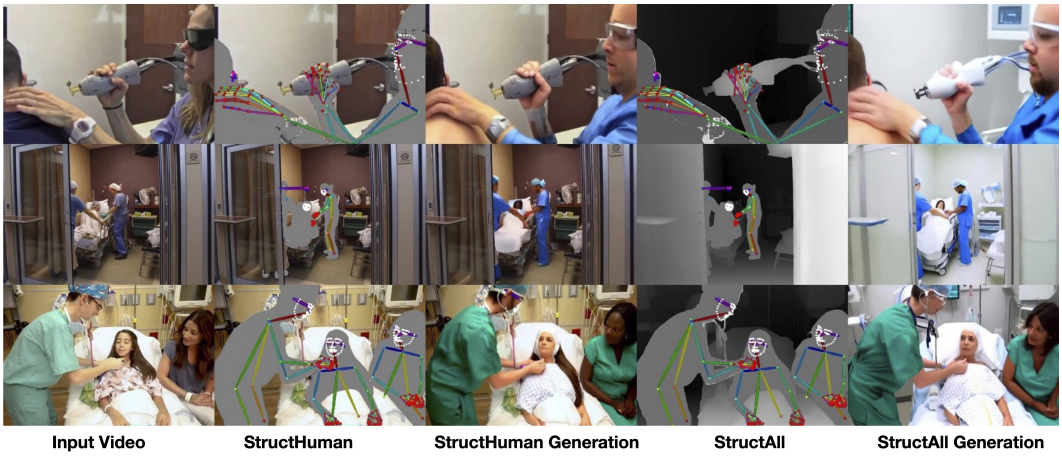

Figure 4: **Medical video anonymization.** Our model is able to produce high-quality anonymizations of clinical videos from MedVideoCap-55K [49] whilst retaining the original contexts. Abridged input captions, top to bottom: *(1)* "A healthcare professional in a blue uniform and protective eyewear applies a handheld laser-like device to a seated patient's back in a clinical room." *(2)* "In a hospital patient room, medical staff in blue scrubs and surgical caps attend to a patient lying in bed under a white blanket." *(3)* "A healthcare professional in green scrubs and a surgical cap listens to a patient's chest with a stethoscope while a family member sits beside the bed."

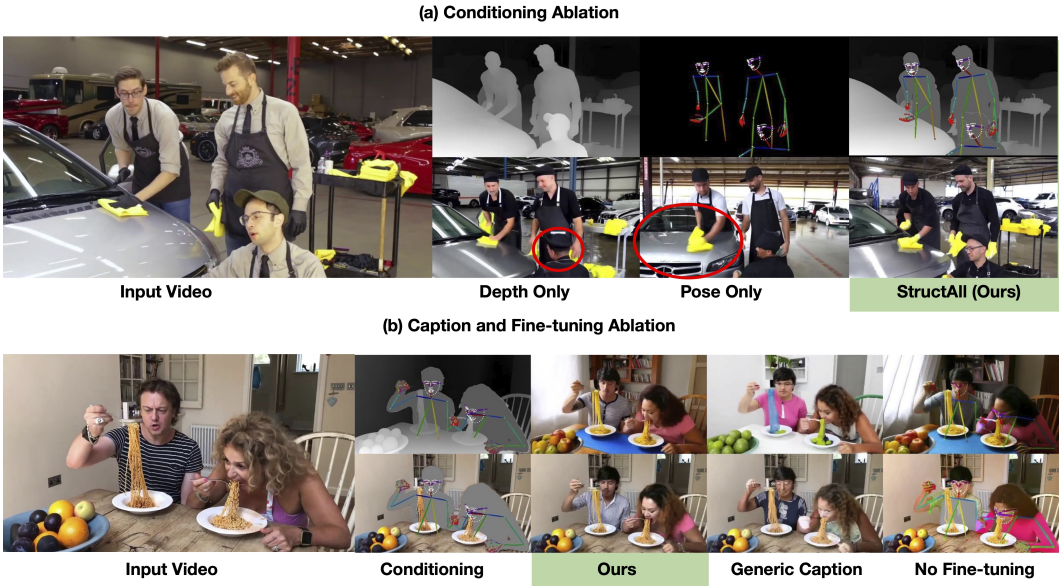

Figure 5: **Ablation study**. We ablate key aspects of our method. Conditioning on depth-only causes pose ambiguity, while pose-only loses scene geometry (*top*). Without a descriptive caption or fine-tuning, generations degrade in quality and become less consistent with the input video (*bottom*).

geometry and results in hallucinations. Using the generic caption "The video depicts one or more humans performing an activity" results in lower quality generations that are inconsistent with the original context. The base model without fine-tuning has trouble handling our layered pose and depth conditioning. Combining all of the components of our method is required for optimal performance.

We conduct further ablations, including swapping in a weaker segmenter, corrupting skeleton keypoints, or contradicting the caption, in the appendix.

5 Discussion and Limitations

We present **ReGenHuman**, a full-body video anonymization pipeline that regenerates the subject from identity-free structural controls instead of editing original frames, resulting in realistic and temporally consistent anonymizations. We show through our experiments that **ReGenHuman** achieves privacy by construction while simultaneously producing realistic high-quality videos that preserve downstream utility. Our work reveals future promise for video anonymization with diffusion-based generative methods and sets a baseline for this relatively new task.

Limitations and Future Work. Our work has a few limitations. *(i). Generation time.* A 49-frame clip ($\approx$6 s) takes a median of 155 s (≈ 2.6 minutes) to generate on a single 48G NVIDIA L40S GPU—in line with the generative baselines, but not real-time. It is, however, a one-time offline cost at release time ($\approx$43 GPU-hours per 1,000 clips). We believe that distillation to a few-step autoregressive student with self-forcing [22] is a natural next step, which will also allow for longer video generation. Clip length is bounded by the Wan2.1-VACE context window; longer videos can already be processed in chunks with reference-image chaining, where the last generated frame of one window becomes the reference for the next so the same synthetic identity carries across the video. Our pipeline is backbone-agnostic and inherits the fidelity of stronger video diffusion models as they are released. *(ii). Motion and silhouette fidelity*. We currently re-use the source pose verbatim, which preserves gait, a known body-level identifier [50]. Perturbing pose sequences and human silhouettes would further extend the privacy guarantee. *(iii). Prompt reliance.* As Wan2.1 is a text-to-video model, generations rely on the prompt, so captioning hallucinations can shift video context (App. B).

Broader impacts. Our work directly enables large-scale release of human-centric video in sensitive areas. Ethical data sharing on a large scale will improve video understanding and generation capabilities of future models, in addition to increasing user trust. Our architectural guarantee also makes the privacy claim auditable: a human can verify that the conditioning inputs contain no private information. Importantly, downstream users should treat generated identities from our method as synthetic, and further work is needed to see if they are provably untraceable to identities from the datasets used to train our model. See App. E for further discussion.

Acknowledgments and Disclosure of Funding

This research was supported in part by the National Institutes of Health grant numbers AG089169, AG102665, and AG084471. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The project was also supported by a pilot award from the Stanford Center for Digital Health.

References

[1] Ahmedt-Aristizabal, D., Armin, M.A., Hayder, Z., Garcia-Cairasco, N., Petersson, L., Fookes, C., Denman, S., McGonigal, A.: Deep learning approaches for seizure video analysis: A review. Epilepsy & Behavior **154**, 109735 (2024)

[2] Annas, G.J.: Hipaa regulations—a new era of medical-record privacy? (2003)

[3] Aslam, N., Ray, A., Haurum, J.B., Esterle, L., Nasrollahi, K.: From pixels to privacy: Temporally consistent video anonymization via token pruning for privacy preserving action recognition. arXiv preprint arXiv:2603.26336 (2026)

[4] Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)

[5] Ban, B., Lee, H.: Faker: Full-body anonymization with human keypoint extraction for real-time video deidentification. arXiv preprint arXiv:2408.11829 (2024)

[6] Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)

[7] Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 961–970 (2015)

[8] Cai, T., Ni, H., Ma, W., Xue, Y., Ma, Q., Leicht, R., Wong, K., Volpi, J., Wong, S.T., Wang, J.Z., et al.: Safetriage: facial video de-identification for privacy-preserving stroke triage. In: International Conference on Information Processing in Medical Imaging. pp. 390–404. Springer (2025)

[9] Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)

[10] Chen, S., Guo, H., Zhu, S., Zhang, F., Huang, Z., Feng, J., Kang, B.: Video depth anything: Consistent depth estimation for super-long videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22831–22840 (2025)

[11] Dai, W., Adeli, E., Luo, Z., Dash, D., Lakshmikanth, S., Durante, Z., Tang, P., Kaushal, A., Milstein, A., Fei-Fei, L., et al.: Developing icu clinical behavioral atlas using ambient intelligence and computer vision. NEJM AI **2**(2), AIoa2400590 (2025)

[12] Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J.S., Salehi, M., Muennighoff, N., Lo, K., Soldaini, L., et al.: Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 91–104 (2025)

[13] Diba, A., Fayyaz, M., Sharma, V., Paluri, M., Gall, J., Stiefelhagen, R., Van Gool, L.: Large scale holistic video understanding. In: European Conference on Computer Vision. pp. 593–610. Springer (2020)

[14] Dinh, M., Jin, S.: Unsafe2safe: Controllable image anonymization for downstream utility. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3326–3336 (2026)

[15] Gawande, U., Hajari, K., Golhar, Y.: Pedestrian detection and tracking in video surveillance system: issues, comprehensive review, and challenges. Recent Trends in Computational Intelligence pp. 1–24 (2020)

[16] Gu, X., Chang, H., Ma, B., Bai, S., Shan, S., Chen, X.: Clothes-changing person re-identification with rgb modality only. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1060–1069 (2022)

[17] HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al.: Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2024)

[18] Hassan, F., Sánchez, D., Soria-Comas, J., Domingo-Ferrer, J.: Automatic anonymization of textual documents: detecting sensitive information via word embeddings. In: 2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). pp. 358–365. IEEE (2019)

[19] He, Y., Huang, F., Jiang, X., Nie, Y., Wang, M., Wang, J., Chen, H.: Foundation model for advancing healthcare: challenges, opportunities and future directions. IEEE Reviews in Biomedical Engineering **18**, 172–191 (2024)

[20] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)

[21] Hu, P., Wang, Z., Sun, R., Wang, H., Xue, M.: M^4I: Multi-modal models membership inference. Advances in Neural Information Processing Systems **35**, 1867–1882 (2022)

[22] Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009 (2025)

[23] Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)

[24] Hukkelås, H., Lindseth, F.: Deepprivacy2: Towards realistic full-body anonymization. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1329–1338 (2023)

[25] Hukkelås, H., Lindseth, F.: Does image anonymization impact computer vision training? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 140–150 (2023)

[26] Hukkelås, H., Mester, R., Lindseth, F.: Deepprivacy: A generative adversarial network for face anonymization. In: International symposium on visual computing. pp. 565–578. Springer (2019)

[27] Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17191–17202 (2025)

[28] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)

[29] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)

[30] Klemp, M., Rösch, K., Wagner, R., Quehl, J., Lauer, M.: Ldfa: Latent diffusion face anonymization for self-driving applications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3199–3205 (2023)

[31] Kung, H.W., Varanka, T., Saha, S., Sim, T., Sebe, N.: Face anonymization made simple. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1040–1050. IEEE (2025)

[32] Li, M., Xu, X., Fan, H., Zhou, P., Liu, J., Liu, J.W., Li, J., Keppo, J., Shou, M.Z., Yan, S.: Stprivacy: Spatio-temporal privacy-preserving action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5106–5115 (2023)

[33] Li, Q., Yu, R., Wang, X.: Vid-sme: Membership inference attacks against large video understanding models. arXiv preprint arXiv:2506.03179 (2025)

[34] Liu, K., Liu, Q., Liu, X., Li, J., Zhang, Y., Luo, J., He, X., Liu, W.: Hoigen-1m: A large-scale dataset for human-object interaction video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24001–24010 (2025)

[35] Ma, Y., He, Y., Cun, X., Wang, X., Chen, S., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4117–4125 (2024)

[36] Mamede, N., Baptista, J., Dias, F.: Automated anonymization of text documents. In: 2016 IEEE congress on evolutionary computation (CEC). pp. 1287–1294. IEEE (2016)

[37] Mopidevi, S., Jang, K.J., Alasaly, B., Pugh, S., Park, J., Batugo, A., Hwang, S., Eaton, E., Mowery, D.L., Johnson, K.B.: Medviddeid: Protecting privacy in clinical encounter video recordings. Journal of Biomedical Informatics p. 104901 (2025)

[38] Pathak, P., Rawat, Y.S.: Colors see colors ignore: Clothes changing reid with color disentanglement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16797–16807 (2025)

[39] Pilán, I., Lison, P., Øvrelid, L., Papadopoulou, A., Sánchez, D., Batet, M.: The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. Computational Linguistics **48**(4), 1053–1101 (2022)

[40] Raina, N., Somasundaram, G., Zheng, K., Miglani, S., Saarinen, S., Meissner, J., Schwesinger, M., Pesqueira, L., Prasad, I., Miller, E., et al.: Egoblur: Responsible innovation in aria. arXiv preprint arXiv:2308.13093 (2023)

[41] Rosberg, F., Aksoy, E.E., Englund, C., Alonso-Fernandez, F.: Fiva: facial image and video anonymization and anonymization defense. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 362–371 (2023)

[42] Sharma, V., Gupta, M., Kumar, A., Mishra, D.: Edunet: a new video dataset for understanding human activity in the classroom environment. Sensors **21**(17), 5699 (2021)

[43] Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE symposium on security and privacy (SP). pp. 3–18. IEEE (2017)

[44] Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6479–6488 (2018)

[45] Vijeikis, R., Raudonis, V., Dervinis, G.: Efficient violence detection in surveillance. Sensors **22**(6), 2216 (2022)

[46] Voigt, P., Von dem Bussche, A.: The eu general data protection regulation (gdpr). A practical guide, 1st ed., Cham: Springer International Publishing **10**(3152676), 10–5555 (2017)

[47] Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)

[48] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)

[49] Wang, R., Chen, J., Ji, K., Cai, Z., Chen, S., Yang, Y., Wang, B.: Medgen: Unlocking medical video generation by scaling granularly-annotated medical videos. arXiv preprint arXiv:2507.05675 (2025)

[50] Weiß, S., Bonenberger, C., Niedermaier, T., Knof, M., Schneider, M.: Privacy beyond the face: Assessing gait privacy through realistic anonymization in industrial monitoring. Sensors **26**(1), 187 (2025)

[51] Xi, D., Wang, J., Liang, Y., Qiu, X., Huo, Y., Wang, R., Zhang, C., Li, X.: Omnivdiff: Omni controllable video diffusion for generation and understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 10915–10923 (2026)

[52] Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9777–9786 (2021)

[53] Xing, J., Xia, M., Liu, Y., Zhang, Y., Zhang, Y., He, Y., Liu, H., Chen, H., Cun, X., Wang, X., et al.: Make-your-video: Customized video generation using textual and structural guidance. IEEE transactions on visualization and computer graphics **31**(2), 1526–1541 (2024)

[54] Xu, L., Zhao, Y., Zhou, D., Lin, Z., Ng, S.K., Feng, J.: Pllava: Parameter-free llava extension from images to videos for video dense captioning. arXiv preprint arXiv:2404.16994 (2024)

[55] Yang, C., Wang, Z., Zhu, X., Huang, C., Shi, J., Lin, D.: Pose guided human video generation. In: Proceedings of the European conference on computer vision (ECCV). pp. 201–216 (2018)

[56] Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4210–4220 (2023)

[57] Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)

[58] Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person re-identification: A survey and outlook. IEEE transactions on pattern analysis and machine intelligence **44**(6), 2872–2893 (2021)

[59] Zhang, J., Ye, M., Yang, Y.: Learnable privacy-preserving anonymization for pedestrian images. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 7300–7308 (2022)

[60] Zhang, X., Zhao, Z.: More effort is needed to protect pedestrian privacy in the era of ai. In: The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track (2025)

[61] Zhang, Y., Qiao, Y., Fricker, J.D.: Investigating pedestrian waiting time at semi-controlled crossing locations: Application of multi-state models for recurrent events analysis. Accident Analysis & Prevention **137**, 105437 (2020)

[62] Zheng, D., Huang, Z., Liu, H., Zou, K., He, Y., Zhang, F., Gu, L., Zhang, Y., He, J., Zheng, W.S., et al.: Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. arXiv preprint arXiv:2503.21755 (2025)

[63] Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., Tian, Q.: Person re-identification in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1367–1376 (2017)

[64] Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3702–3712 (2019)

[65] Zhu, B., Fang, H., Sui, Y., Li, L.: Deepfakes for medical video de-identification: Privacy protection and diagnostic information preservation. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 414–420 (2020)

[66] Zwick, P., Roesch, K., Klemp, M., Bringmann, O.: Context-aware full body anonymization using text-to-image diffusion models. arXiv preprint arXiv:2410.08551 (2024)

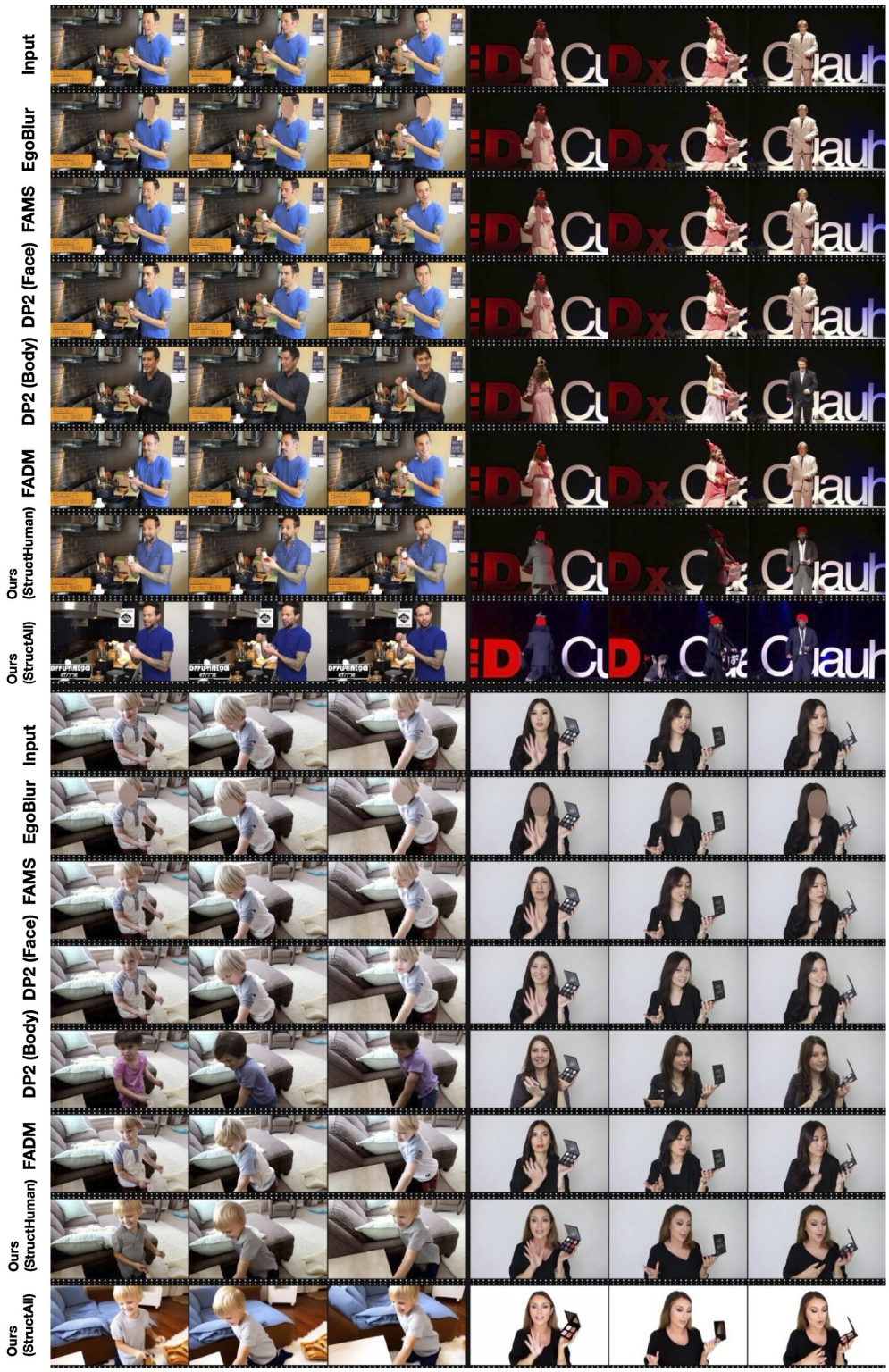

Figure 6: Qualitative comparison on 4 randomly sampled examples from our test subset of HOIGen-1M [34]. EgoBlur [40], FAMS [31], and DeepPrivacy2 (Face) [24] do not anonymize the body, resulting in less privacy. DeepPrivacy2 (Body) [24] and FADM [66]'s anonymizations are temporally inconsistent. Our method consistently produces high-quality, temporally consistent anonymizations that can anonymize either just the human or both the human and background.

A Qualitative Results

We provide additional qualitative results in Figure 6 and as a video in the supplementary materials.

B Additional Studies

B.1 Human realism study

To complement the automated Realism metric, we ran a user study: 17 raters each rated 15 clips (3 scenes $\times$ 5 conditions: real, DeepPrivacy2-body, FADM, StructHuman, StructAll) in randomized order with methods unnamed, on a 1–5 scale (1 = "obviously fake", 5 = "looks real").

Condition	MOS (1–5) ↑	95% CI (over raters)
Real video (hidden reference)	4.47	[4.18, 4.76]
DeepPrivacy2 (body) [24]	1.10	[1.00, 1.20]
FADM [66]	1.43	[1.18, 1.69]
StructHuman (ours)	**2.24**	[1.69, 2.78]
StructAll (ours)	**2.49**	[2.08, 2.90]

Table 5: **Human realism study** (17 raters $\times$ 15 clips, randomized and unnamed).

All four pre-registered ours-vs-generative-baseline rater-paired contrasts are significant after Holm correction: StructAll$-$FADM $+1.06$ and StructAll$-$DP2-body $+1.39$ (both $p < 10^{-4}$, 16/17 raters), StructHuman$-$FADM $+0.80$ ($p = 0.0015$, 13/17), StructHuman$-$DP2-body $+1.14$ ($p = 0.0004$, 13/17); StructAll vs. StructHuman is not separable ($p = 0.21$). Inter-rater reliability: Krippendorff's α (ordinal) $= 0.65$ and ICC(2,k) of the reported means $= 0.98$ (α is deflated by scale-floor compression; the high ICC confirms the means are stable). The hidden real reference (4.47) shows raters readily detect generated content in general; the supported claim is that our outputs are substantially *more* realistic than all prior generative body anonymizers, not that they pass as real.

B.2 Conditioning faithfulness and robustness

We probe how the model responds when each conditioning input is deliberately corrupted, on the same 100-clip seeded subset with the paper checkpoint and evaluation suite. Across all three experiments the response is *faithful and localized*: corrupting one conditioning input never changes privacy significantly.

Segmenter swap. Rebuilding the entire StructAll conditioning stack with a deliberately weaker, older segmenter (Grounded-SAM in place of SAM3) and regenerating all clips changes identity similarity by 0.0001, text similarity by 0.0162, and realism by 0.41 points — privacy does not degrade and no quality metric moves by more than half a point.

Skeleton corruption. We corrupt the extracted skeletons between detection and rendering (Gaussian jitter on every keypoint plus random keypoint dropout, seeded), re-render, rebuild the conditioning, and regenerate:

StructAll (100 clips)	Identity Sim. ↓	Text Sim. ↓	Realism ↑	Human Anatomy ↑	Subject Cons. ↑	Motion Smooth. ↑
Clean DWPose (paper)	0.0034	0.0768	89.85%	88.33%	93.20%	96.79%
Jitter σ=1%H + 5% dropout	0.0038	0.0852	90.55%	87.94%	93.26%	96.68%
Jitter σ=3%H + 20% dropout	0.0030	0.1010	90.00%	85.63%	93.05%	96.52%

Table 6: **Skeleton-corruption ablation.** Even severe corruption (every keypoint displaced by 3% of image height, one in five keypoints deleted) costs only 2.7 points of Human Anatomy — the axis skeletons govern — with privacy invariant.

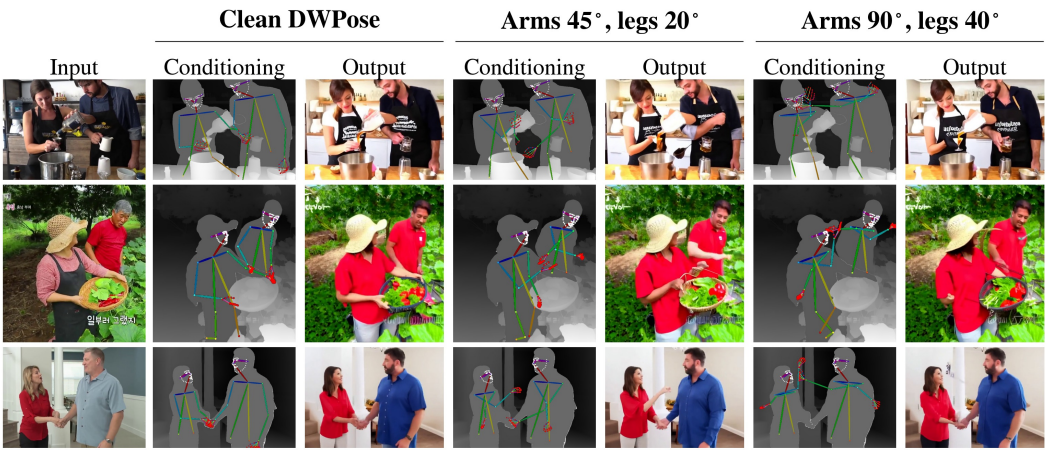

Figure 7: **Effect of ill-conditioned pose on the output** (StructAll; same frame of each clip). Going beyond the keypoint jitter and dropout of Table 6, we physically displace every person's limbs — each arm rigidly rotated about its shoulder and each leg about its hip by a fixed angle for the whole clip — and regenerate with the same checkpoint. At moderate displacement the generated people follow the moved limbs (the woman's raised arm in the handshake, a different grip on the basket, the man's hand placement while pouring). At severe displacement, limbs pushed far outside a person's silhouette or to infeasible positions are overridden by the model's priors on human pose.

Caption contradiction. We rewrite captions to contradict the conditioning (replacing all plural references to people with singular ones). Realism drops by 1.4 points and anatomy drops by 0.8 points, showing that the prompt mainly guides appearance while the structural conditioning supplies the geometry.

StructAll (n=43)	Identity Sim. ↓	Text Sim. ↓	Realism ↑	Human Anatomy ↑	Subject Cons. ↑	Temporal Flicker. ↑	Aesthetic Quality ↑
Original caption	0.0039	0.0533	89.77%	89.63%	93.49%	93.92%	53.76%
De-pluralized caption	0.0035	0.0362	88.37%	88.84%	93.58%	93.93%	53.45%
Δ	−0.0004	−0.0171	−1.40%	−0.79%	+0.09%	+0.01%	−0.31%

Table 7: **Caption-contradiction ablation.** Captions rewritten to describe a single person while the conditioning still depicts the original multi-person scene, evaluated on the clips whose captions the rewrite actually altered. Privacy does not degrade (both similarities decrease) and the quality cost is at most 1.4 points.

B.3 LoRA capacity

We train additional StructAll adapters at rank ρ=16 and ρ=64 and compare at a matched step (8,000). It is clear that behavior depends on our structural conditioning rather than adapter capacity, and ρ=32 is a reliable standard default.

LoRA rank (StructAll)	Identity Sim. ↓	Text Sim. ↓	Realism ↑	Anatomy ↑
$\rho = 16$	0.0036	0.0926	90.25%	88.88%
$\rho = 32$ (paper)	0.0037	0.0883	90.45%	90.00%
$\rho = 64$	0.0039	0.1001	89.45%	88.86%
Spread (max − min)	0.0003	0.0118	1.00	1.14

Table 8: **LoRA rank ablation.** No privacy metric moves by more than 0.0003 and no quality metric by more than 1.2 points.



Figure 8: **Study of prompt reliance.** Our pipeline relies heavily on input captions. Here we show generations resulting from modifying the demographic of the input caption from left to right of "a young woman" (the default) to "a young man" and "an elderly woman". The model generates a similar background and foreground in all cases, but seamlessly swaps out the person.



Figure 9: **Failure cases.** Here we show some failure cases of our model on MedVideoCap-55K [49]. On the leftmost image, the foot has numerous artifacts and is ambiguously wrapped in a cast. In the next image, a nurse is shown simultaneously putting on a mask and wearing one. The model is not able to generate a doctor wearing a stethoscope, and struggle with generating IVs as well.

B.4 Runtime

Median measured anonymization time per 49-frame clip on a single NVIDIA L40S: DeepPrivacy2 (body) 12 s, FAMS 88 s, FADM 122 s, **StructAll (ours) 155 s**. Our runtime matches the generative baselines; the GAN-based DeepPrivacy2 is faster but rates at the floor of the human study (App. B.1). Anonymization is an offline, one-time operation performed at dataset-curation or release time: at $\sim$2.6 min/clip, a 1,000-clip release costs roughly 43 GPU-hours, and our pipeline inherits any few-step distillation speedup of the backbone.

B.5 Prompt reliance and failure cases

We further study our pipeline's reliance on the input caption in Figure 8. By simply editing the demographic in the input caption, we are able to edit the appearance of the generated human. This allows for seamless demographic replacement in addition to anonymization. We note that demographic randomization fundamentally changes the context of the video (i.e. a downstream video question answering task may have differing labels aligned with the input video). Nevertheless, we believe this capability may be promising for certain use cases where demographics are not relevant to the context of the task.

While we believe that our combination of depth, pose, and descriptive text caption provides strong conditioning for many anonymization use cases, our model still inherits many of the failure cases of video diffusion models. We show a few failures we observed on MedVideoCap-55K in Figure 9. The failures largely stem from ambiguous input conditioning and niche contexts. More work is needed to be done to assist video diffusion models in becoming more context-aware.

C Dataset details

We train with a subset of HOIGen-1M [34] (`https://huggingface.co/datasets/HOIGen/HOIGen-1M`), distributed under an Apache-2.0 tag with a gated access condition restricting use to non-commercial purposes. We validate with MedVideoCap-55K [49] (`https://huggingface.co/datasets/FreedomIntelligence/MedVideoCap-55K`; Apache-2.0 tag, released strictly for research use and not for clinical practice), NExT-QA [52] (`https://github.com/doc-doc/next-qa`; MIT-licensed annotations and code, with videos sourced from VidOR under its own terms), and CCVID [16] (released by its authors for research use alongside Apache-2.0-licensed code, `https://github.com/guxinqian/Simple-CCReID`). We use all four datasets for non-commercial research only, in accordance with these terms, and release no data from them.



Figure 10: **Classical anonymization baselines.** Visualization of our implementations of redaction-based anonymization baselines. All approaches use the same SAM3 [9] extractions as our methods.

We utilize a randomly sampled subset of 7,500 clips from HOIGen-1M [34] for training. We use 49-frame clips at 832×480 and 8 fps. We use the same frame sampling regime for our testing on our subsets of CCVID [16], NExT-QA [52], and MedVideoCap-55K [49].

D Implementation Details

D.1 Model training and conditioning extraction

Training. **StructHuman** and **StructAll** are trained as *two separate adapters* in two independent runs, each on its own conditioning; no weights are shared beyond the frozen base model. For each, we train a rank-ρ=32 LoRA (with α=32 and no dropout) on the projection matrices $\{q, k, v, o, \mathrm{ffn}_0, \mathrm{ffn}_2\}$ of every Context Adapter block of a frozen Wan2.1-VACE-1.3B [47, 27]. The Wan DiT backbone and the pretrained VACE Context Adapter stay entirely frozen, and the LoRA factors are trained with the rectified-flow objective of Eq. 4 (Sec. 3). We optimize with AdamW at a constant learning rate 10^{-4}, batch size 1 per GPU (effective batch size 2), in bfloat16 with gradient checkpointing, on 49-frame clips at 832×480. We train each adapter for 16,000 iterations (a single run per variant; no seed averaging — reported numbers aggregate over the 1,000 evaluation clips), taking approximately 72 hours with 2 NVIDIA L40S GPUs. At inference, we perform 50 rectified-flow steps per clip, using a classifier-free guidance scale 5.0. Inference takes a median of 155 s per 49-frame clip on a single NVIDIA L40S GPU (App. B.4). We utilize the training infrastructure of ModelScope's DiffSynth-Studio (`https://github.com/modelscope/diffsynth-studio`), which is licensed under the Apache-2.0 license.

StructHuman mask-coverage mitigation. StructAll's privacy guarantee (Eq. 3) is independent of conditioning quality: an inaccurate mask or skeleton only mis-places where the depth map is flattened or how the pose is drawn, and never creates a path for original pixels to reach the output. For StructHuman, under-segmentation could retain original human pixels in the preserved background, so the guarantee is conditional on mask coverage. We therefore recommend two deployment-time mitigations, which we will ship with our code release: (a) dilating the human mask by a safety margin before compositing, and (b) an automatic leak check that runs a person/face detector over the retained background region, flagging clips for re-masking or routing them to StructAll.

Conditioning extraction. We use SAM3 [9] for human masks (the released `facebook/sam3` checkpoint, text-prompted with "person" on the first frame and propagated through the clip; all detected instances are merged into one binary mask, used raw with no dilation, hole-filling, or temporal smoothing), DWPose [56] for the rendered skeleton (the whole-body `dw-ll_ucoco_384` ONNX model with a YOLOX-L person detector), and Video Depth Anything [10] (ViT-Large) for monocular depth, normalized per-frame by min–max to $[0, 255]$ grayscale. Conditioning is extracted at the source resolution on every 4th frame (≈8 fps). We utilize per-clip captions directly from the datasets for HOIGen-1M and MedVideoCap-55K and prompt Qwen3-VL-8B-Instruct [4] to generate

a detailed caption of the scene when they are not provided (a single dense paragraph covering scene, people, actions over time, and camera framing). All extractors are frozen.

Baselines. For our classical redaction baselines ("Gaussian Blur", "Pixelation", "Blackout"), we take the extracted SAM3 [9] human masks and conduct the corresponding operations on those pixels. See Figure 10 for visualizations.

- **Gaussian blur.** The transformed image is obtained by applying a Gaussian filter to the entire frame:
$$\mathbf{T} = G_{k,\sigma}(\mathbf{I}), \tag{5}$$
where $G_{k,\sigma}$ denotes a Gaussian blur with kernel size $k = 51$ and standard deviation $\sigma = 30.0$. The large kernel and high sigma ensure that faces and fine-grained details (e.g., text on clothing) become unrecognizable, while rough color and shape information is retained.
- **Pixelation.** The transformed image is produced via a two-step resize operation:
$$\mathbf{T} = \texttt{resize}^{\text{NN}}_{\uparrow}\big(\texttt{resize}^{\text{bilinear}}_{\downarrow}(\mathbf{I},\ b \times b)\,,\ H \times W\big)\,, \tag{6}$$
where $b = 30$ is the block size. The frame is first downsampled to a 30×30 grid using bilinear interpolation, then upsampled back to the original resolution using nearest-neighbor interpolation. This produces the characteristic mosaic effect, where each output "pixel" covers approximately $\frac{H}{30} \times \frac{W}{30}$ of the original image. Human regions become coarse color blocks that prevent identification while preserving gross spatial layout.
- **Blackout.** The simplest baseline directly zeroes out all masked pixels:
$$\mathbf{O} = \mathbf{I} \odot (1 - \mathbf{M}), \tag{7}$$
requiring no additional parameters. This constitutes the most aggressive classical approach: all appearance, texture, and color information within human regions is completely removed, leaving only the silhouette boundary as implicit shape information.

For the classical baselines the transformed frame replaces the original only inside the mask, $\mathbf{O} = \mathbf{M} \odot \mathbf{T} + (1 - \mathbf{M}) \odot \mathbf{I}$; the masks are the same raw SAM3 masks our method uses.

D.2 CCVID re-identification attack

We use CSCI-V [38] released checkpoint `ez_eva02_vid_hybrid_extra_best.pth` as our attacker. We set the gallery as clean unanonymized tracklets and query with anonymized tracklets, simulating the release of reference footage. We randomly sample 500 query sequences as our test set; both of our variants in Table 2 use this identical attacker, tracklet subset, and caption configuration.

Caption protocol. CCVID tracklets carry no captions, so we caption each clip with Qwen3-VL and replace the demographic/appearance descriptors with uniformly sampled random ones before generation.

Resolution. CCVID tracklets are pre-cropped 256×128 pedestrian crops. For our method, each crop is scaled aspect-preservingly to fit the 832×480 generation canvas and center-padded with black; after generation, the padding is removed and the frames are resized back to the native crop size before being passed to CSCI-V. All baselines run directly on the native crops with the settings above; when EgoBlur or a face-only method finds no face in a crop, the frame passes through unchanged, which is why face-only methods stay highly re-identifiable in Table 2.

D.3 Zero-Shot Video-QA

Question construction. For each clip we derive three question types from the clip caption: an *action* question ("what activity is the person performing?"), an *object* question ("what objects are the people interacting with?"), and an *interaction* question ("how is the person manipulating the object?"). Multiple choice questions (four options each, one per type, i.e. 3 MCQs per clip) are generated by prompting a frozen Qwen3-VL-8B-Instruct model with the clip's initial frame and its corresponding detailed caption. Ground-truth answers are extracted from the caption with a small



Figure 11: **Sample generated MedVideoCap-55K questions.** The questions challenge the VLM to reason about the anonymized frames, testing if they remain consistent with the original input video.

regex+LLM pipeline; clips for which extraction fails are dropped. The HOIGen evaluation set is the same 1,000-clip subset as Section 4.3, yielding 3,000 questions per method.

Evaluation. A frozen Qwen3-VL-8B [4] is prompted at 448×448, 8 uniformly-sampled frames per clip, with a fixed system prompt. We score model answers against the ground truth with exact-match; the reported Acc. (%) is the mean over questions.

D.4 MedVideoCap protocol

Question construction. We replicate the question construction and evaluation pipeline for the MedVideoCap clips detailed in App. D.3. We provide sample questions and answers in Figure 11.

Subset. A 1,500-clip randomly sampled MedVideoCap-55K [49] subset, split into 1,000 clips of train and 500 clips of test. We filter the dataset to include only "clinical practice videos" whose captions contain mentions of patients and doctors. The resulting dataset only includes full-body clinical context videos, optimal for our video anonymization investigations.

E Broader impacts and misuse

Two models, by design. We propose two variants, **StructHuman** and **StructAll**, rather than a single "one-size-fits-all" anonymizer. The two condition on different amounts of scene structure — **StructHuman** preserves only the human pose and mask, while **StructAll** additionally preserves a coarse depth backdrop — and they trade off privacy and utility differently. Forcing a single model would either over-anonymize footage where the surrounding context is itself the signal of interest (e.g. a clinical room layout or relevant clinical equipment), or under-anonymize footage where the background carries identifying cues (license plates, signage, identifiable interiors). We allow the deployer match the anonymizer to the downstream task and threat model.

Not a deepfake tool, but not risk-free. Our pipeline is deliberately built without a ground-truth reference frame of any specific person: the conditioning signals (pose, human mask, optional depth) are identity-agnostic, and the model is trained to *break* identity rather than reproduce it. This is a meaningfully different design point from face-swap or talking-head systems that take a target identity as input. That said, the text prompt is still an open channel: a user can describe a specific person — by name, profession, demographic, or distinctive attribute — and the underlying video diffusion prior may steer the generated human toward that description. We do not consider this a failure of the anonymization objective (the original identity in the source video is still removed), but it is a possible misuse inherited from the base text-to-video model. In our own pipeline the caption is produced automatically and appearance descriptors are randomized before generation (App. D.2), so the prompt is never a free-form identity description. For deployments that accept user prompts, we recommend applying the same sanitization: filtering named individuals and identity-specific descriptors from the prompt before generation, and logging prompts alongside released outputs so that misuse can be audited.

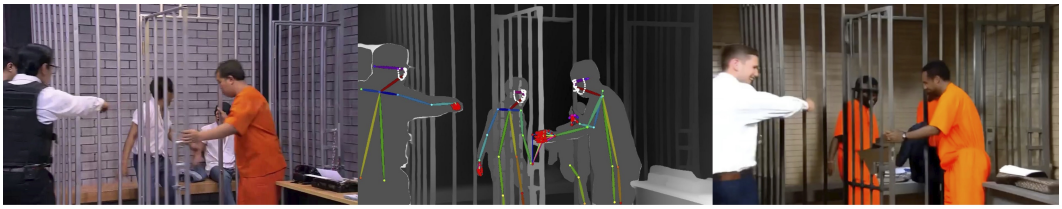

Figure 12: **Racial bias.** We show one example of bias in our anonymization pipeline. The video generation model assumes Black males are the incarcerated subjects and a White male is the warden.

Training data and memorization. The datasets used in this work are released for non-commercial research use (App. C lists each dataset's terms), and we use them only under those terms. However, like any large generative model fine-tuned on video, our LoRA may to some extent memorize training distributions, and a small fraction of generated humans may resemble individuals in the training set. We did not observe verbatim identity reproduction in our qualitative inspection, but we cannot rule it out at scale, and we recommend that downstream users treat anonymized outputs as "de-identified" instead of fully new.

Demographic bias. The model inherits demographic biases from both Wan2.1-VACE's pretraining and our fine-tuning data. One such case is evident when prompted to generate a prison setting — the model generates Black male subjects as prisoners and white men as wardens, even when the source prompt never explicitly mentions race (Fig. 12). We attribute this primarily to biases in the base video generation model's training distribution rather than to our approach. One practical mitigation for this task is to inject explicit demographic randomization into the prompt (e.g. sample a race/gender token uniformly from a vocabulary before generation — see Figure 8).

Net assessment. The capabilities our system exposes — text-and-structure-conditioned video synthesis of humans — already exist in publicly released video-to-video and text-to-video models. Our contribution is a recipe that bends those capabilities toward privacy preservation rather than identity reproduction, and our threat model (Sec. D.2) explicitly evaluates the privacy gain rather than asserting it. We therefore believe the marginal misuse risk introduced by this work is small relative to its benefit for downstream privacy-respecting computer vision. We encourage deployers to (a) keep the source footage out of public release, (b) apply prompt-side filtering against named individuals and protected attributes, and (c) prefer **StructAll** over **StructHuman** when the surrounding scene is itself sensitive.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the contributions of the work. The claims made match the results presented in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide limitations in Section 5, as well as further discussion in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Detailed implementation details are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We provide detailed implementation details in the appendix, but do not have immediate plans to release the code.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Detailed implementation details are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The metrics we utilize (typically used to benchmark video generation models) are typically not reported with error bars.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide details on our compute resources in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and believe our work conforms to the guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss broader impacts of our work in Section 5 and in Appendix E.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The pre-trained model we use is open-source, and the datasets we train on are licensed with the Apache 2.0 license, so we believe no risk of misuse exists.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the code package and datasets we use in the appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Our submission does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.