MEANINGFUL ANONYMIZATION FOR PRIVACY-PRESERVING VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Videos contain rich visual and temporal information for understanding human activities, but can also expose sensitive information about people and their surroundings. Existing privacy-preserving video methods primarily reduce such leakage by suppressing or discarding privacy-sensitive visual information, which can inadvertently remove content useful for understanding the video. We introduce **Mask2Safe**, a privacy-preserving video generation framework that learns *where* privacy-preserving modifications are needed and uses this localization to guide *how* sensitive content is transformed. Rather than relying on the costly vision-language reasoning for video frames, Mask2Safe learns semantic privacy localization from paired unsafe–safe images through a lightweight *Privacy Masker*. Its predictions are refined and propagated over the whole clip to produce temporally consistent masks that support either mask-based or generative anonymization. We further introduce **SafeQA**, a video question-answering benchmark that evaluates meaningful anonymization along three complementary dimensions: *Privacy*, whether the original sensitive information remains recoverable; *Visibility*, whether the affected visual content remains observable; and *Local Utility*, whether nearby non-sensitive information remains recoverable. Experiments on VP-UCF101 and VP-HMDB51 show that Mask2Safe reduces the predictability of sensitive attributes while retaining downstream utility on action recognition task, and that classifiers trained on our anonymized videos are able to transfer effectively to the original video domain. SafeQA further shows that generative anonymization substantially improves visibility over mask-based anonymization while maintaining strong local utility.

1 INTRODUCTION

Video provides rich information about actions, interactions, and temporal dynamics, but this richness also makes it particularly privacy-sensitive: identity, demographic attributes, personal objects, and other sensitive information may persist across frames and become encoded in learned representations. Privacy-preserving video learning therefore seeks to reduce privacy leakage while retaining information useful for downstream tasks (Ryoo et al., 2017; Ren et al., 2018).

Existing methods primarily approach this trade-off by learning what information can be *suppressed*. SPAct (Dave et al., 2022) learns an input anonymizer that disrupts privacy-relevant representations while preserving action recognition; TeD-SPAD (Fioresi et al., 2023) incorporates temporal distinctiveness for privacy-preserving video anomaly detection; and more recent methods selectively anonymize or prune spatiotemporal representations (Li et al., 2023; Aslam et al., 2026). Despite their different mechanisms, these approaches largely improve privacy by transforming, suppressing, or discarding privacy-bearing visual information.

Privacy-sensitive information, however, often overlaps with visual content useful for understanding the video. Removing or obscuring a person or object may protect private attributes while simultaneously discarding appearance, spatial, or contextual cues. We therefore ask a different question: *rather than determining how much sensitive information can be removed, can we replace it with privacy-safe content while preserving the remaining visual semantics?* Unsafe2Safe (Dinh & Jin, 2026) recently explored this direction for images by using vision-language reasoning and generative editing to replace sensitive content with realistic privacy-safe alternatives.

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

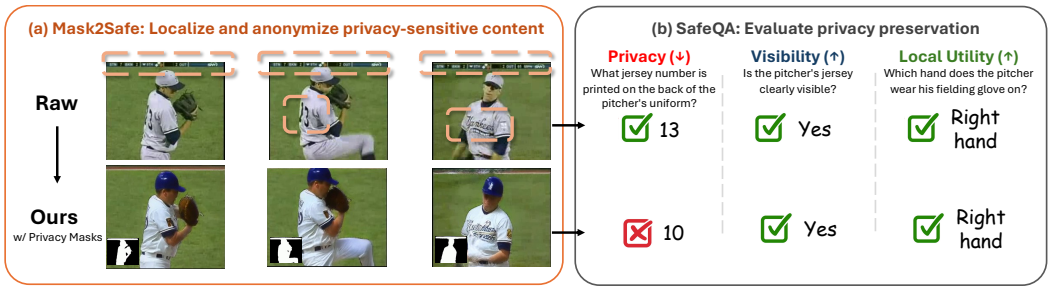

Figure 1: **Meaningful privacy-preserving video generation and evaluation.** (a) Mask2Safe localizes privacy-sensitive regions across video frames and replaces them with plausible, privacy-safe content while preserving useful visual information. (b) SafeQA separately evaluates recovery of the original sensitive information, visibility of the affected content, and preservation of nearby non-sensitive information. Additional qualitative examples are provided in Appendix C.2.

Extending this idea to video introduces a key challenge: privacy decisions must remain consistent over time. Sensitive content may move, become occluded, or reappear, while independently applying image-level privacy reasoning and generation to every frame is computationally expensive and prone to inconsistent localization and editing. This motivates our central question: *can semantic privacy knowledge learned from images be transferred to video to localize sensitive content over time and guide temporally consistent privacy-preserving transformations?*

We introduce **Mask2Safe**, a privacy-preserving video generation framework that transfers privacy knowledge learned from images to video. Using paired unsafe–safe images, we train a lightweight **Privacy Masker** to learn which visual regions require privacy-preserving modification without repeated frame-wise VLM reasoning. At inference time, its predictions are refined and propagated over time using promptable video segmentation (Ravi et al., 2025), producing temporally consistent masks that support either mask-based anonymization or generative anonymization with plausible privacy-safe alternatives. This design separates learning *where* privacy-preserving changes are required from *how* those regions are transformed.

Evaluating such transformations also requires going beyond conventional privacy prediction and downstream action recognition, which do not reveal whether anonymization preserves the non-sensitive visual information associated with affected regions. We therefore introduce **SafeQA**, a video question-answering benchmark that separately evaluates whether the original sensitive information remains recoverable, whether the affected visual content remains observable, and whether nearby non-sensitive information remains recoverable.

We evaluate Mask2Safe on VP-UCF101 and VP-HMDB51 (Li et al., 2023) using privacy prediction, action recognition, anonymized-to-original domain transfer, and SafeQA. Our results show that Mask2Safe reduces sensitive-attribute predictability while retaining action information, while classifiers trained on our anonymized videos transfer effectively to the original video domain. SafeQA further shows that generative anonymization substantially improves visibility over mask-based anonymization while maintaining strong local utility. Hence, these results demonstrate the potential of transferring semantic privacy knowledge from images to video and replacing localized sensitive content in lieu of broadly suppressing useful visual information.

Our contributions are as follows:

- We introduce **Mask2Safe**, a privacy-preserving video generation framework with a lightweight **Privacy Masker** that learns semantic privacy localization from paired unsafe–safe images and transfers this knowledge to video without repeated frame-wise VLM reasoning. The localized regions are temporally refined and used for either mask-based or generative anonymization.
- We introduce **SafeQA**, a video question-answering benchmark that evaluates meaningful anonymization along three complementary dimensions: *Privacy*, measuring whether the original sensitive information remains recoverable; *Visibility*, measuring whether the affected visual content remains observable; and *Local Utility*, measuring whether nearby non-sensitive information remains recoverable.

- We demonstrate on VP-UCF101, VP-HMDB51, and SafeQA that Mask2Safe reduces privacy leakage while retaining action-relevant and fine-grained non-sensitive information, and that models trained on the resulting anonymized videos transfer effectively to the original video domain.

2 Related Work

2.1 Privacy-Preserving Video Understanding

Privacy-preserving video understanding typically balances privacy and utility by limiting sensitive visual information available to downstream models. Prior approaches include fidelity degradation (Ryoo et al., 2017), motion-based representations (Kumawat & Nagahara, 2022), optical transformations (Hinojosa et al., 2022; Boloor et al., 2025), coded-aperture systems (Wang et al., 2019), face obfuscation (Tomei et al., 2021), and encrypted representations (Ishikawa et al., 2024; Chen et al., 2026). Other methods jointly optimize privacy and downstream utility (Dave et al., 2022; Li et al., 2023; Wu et al., 2022; 2018; Aslam & Nasrollahi, 2025), including learned anonymization (Dave et al., 2022; Fioresi et al., 2023) and selective suppression of privacy-sensitive video representations (Li et al., 2023; Aslam et al., 2026). Because privacy is obtained by degrading or discarding the affected content, non-sensitive information co-located with sensitive regions is typically lost as well.

Selective and generative anonymization instead localizes sensitive content and modifies only the affected regions, using noise (Ilic et al., 2024) or synthetic replacements (Maximov et al., 2020; Schneider et al., 2024; Azhar et al., 2025; Egin et al., 2025). Unsafe2Safe (Dinh & Jin, 2026) extends this direction beyond the closed-set faces and bodies to an open set of sensitive attributes in images, but relies on VLM reasoning and editing individually for each image. Rather than applying this pipeline to every frame, Mask2Safe distills its privacy localization from unsafe–safe image pairs into a lightweight Privacy Masker and relies on video segmentation and editing to keep the anonymization temporally consistent.

Existing video anonymizers are typically learned from videos in the downstream distribution (Dave et al., 2022; Fioresi et al., 2023; Li et al., 2023); even cross-domain approaches may require unlabeled target-domain videos (Xia et al., 2025) or an anonymization module at downstream inference (Fioresi et al., 2026). In contrast, our Privacy Masker is learned entirely from external image transformations and requires neither downstream training videos nor the Privacy Masker at downstream inference. Mask2Safe produces standard RGB videos that can be used directly by downstream video models.

2.2 Evaluating Privacy-Preserving Video Systems

Privacy-preserving video systems are commonly evaluated through a privacy–utility trade-off, with privacy measured by predicting predefined attributes and utility by downstream task performance. PA-HMDB51 (Wu et al., 2022) and VP-HMDB51/VP-UCF101 (Li et al., 2023) evaluate predefined privacy attributes, while other work considers broader privacy analyses (Fioresi et al., 2026) or identity leakage through re-identification (Thapar et al., 2021; Moon et al., 2023; Sun et al., 2026). These evaluations measure whether sensitive attributes or identities remain recoverable, but because utility is scored against separate, clip-level targets, they do not characterize what happens to the content inside the anonymized regions.

Recently, video question answering provides a fine-grained mean to assess retained information (Xiao et al., 2021; Li et al., 2024; Fu et al., 2025; Yuan et al., 2025). ReGenHuman (Sun et al., 2026), for example, uses caption-derived questions about activities, objects, and interactions to evaluate utility separately from privacy. On the contrary, SafeQA anchors both privacy and preservation questions to the *same privacy-relevant content*, so that protection and retained information are measured on the regions that were actually modified.

3 Mask2Safe: Privacy-Preserving Video Generation

Our goal is to transform a video so that privacy-sensitive visual information is removed or replaced while preserving the remaining visual content. A central challenge is determining *where* privacy-

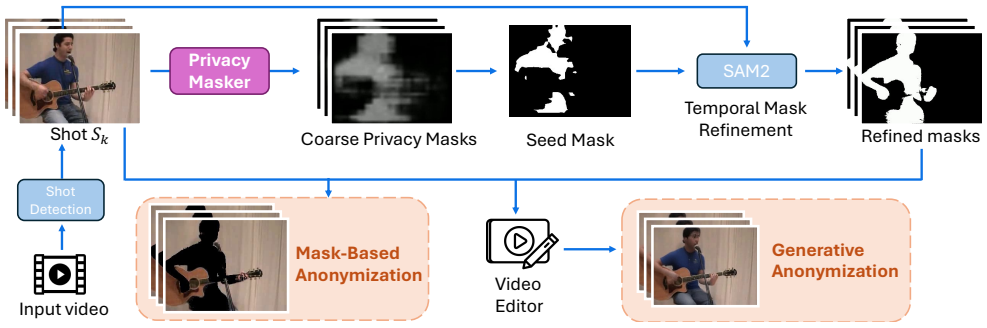

Figure 2: **Mask2Safe video anonymization pipeline.** Frame-wise Privacy Masker predictions are refined and propagated within each shot using SAM2, producing temporally consistent masks for either mask-based or generative anonymization.

preserving modifications should be applied throughout a video. Learning this localization directly from paired unsafe and privacy-preserving videos would require corresponding video pairs in which sensitive content is consistently modified across space and time, which are difficult to obtain at scale. Instead, we transfer privacy knowledge learned from image-level transformations to video.

Our key observation is that paired unsafe and privacy-preserving images provide implicit supervision for privacy localization: regions that must change to reproduce a privacy-preserving counterpart indicate where privacy-sensitive information resides. We leverage source–target image pairs from Unsafe2Safe (Dinh & Jin, 2026) to train a lightweight *Privacy Masker* that learns this localization without manually annotated privacy masks or paired privacy-preserving videos.

Figure 2 provides an overview of Mask2Safe. Given an input video, the Privacy Masker identifies coarse privacy-sensitive regions in individual frames. We perform shot-aware seed selection and use SAM2 (Ravi et al., 2025) to refine and propagate these regions, producing spatially precise and temporally consistent privacy masks. The resulting masks can either directly suppress sensitive regions or guide a generative video editor to replace their content with plausible alternatives.

3.1 Learning Privacy Localization

We train the Privacy Masker using source–target image pairs (I_s, I_t) from Unsafe2Safe (Dinh & Jin, 2026), where I_t is a privacy-preserving edit of the source I_s, or is identical to I_s when no modification is required. As illustrated in Fig. 3, rather than supervising the masker with annotated privacy regions, we learn localization through the source–target transformation: the predicted mask should identify where changes are necessary to reproduce the privacy-preserving target while leaving the remaining content unchanged.

We encode the source and target images into the latent space of a frozen Stable Diffusion v1.5 VAE (Rombach et al., 2022),

$$z_s = \mathcal{E}(I_s), \qquad z_t = \mathcal{E}(I_t), \tag{1}$$

and extract source features $F_s = \Phi(I_s)$ using a frozen DINOv2 encoder (Oquab et al., 2024). A trainable mask decoder f_θ predicts a soft privacy mask

$$m = \sigma(f_\theta(F_s)), \qquad m \in [0,1]^{H_z \times W_z}. \tag{2}$$

Because ground-truth privacy masks are unavailable, we learn m through its role in reconstructing the privacy-preserving target. A text-free U-Net G predicts a candidate replacement latent r, which is composited with the source according to

$$\hat{z} = m \odot r + (1 - m) \odot z_s. \tag{3}$$

where $\odot$ denotes element-wise multiplication. Thus, m determines *where* the source should change, while r determines *what* replaces the corresponding content.

We supervise the composited latent against the privacy-preserving target and additionally constrain the replacement to preserve source content outside the predicted region and favor the privacy-preserving target within it. In addition, an area regularizer discourages unnecessarily large masks.

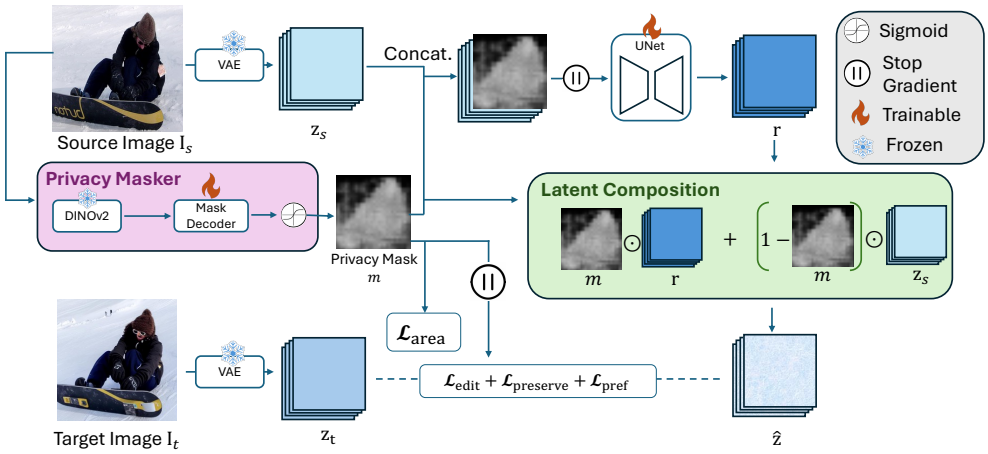

Figure 3: **Learning the Privacy Masker.** The mask m localizes regions needed to reproduce the privacy-preserving target, learned from unsafe–safe image pairs without mask annotations.

The complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{edit}} + \mathcal{L}_{\text{preserve}} + \mathcal{L}_{\text{pref}} + \lambda_{\text{area}}\mathcal{L}_{\text{area}}, \tag{4}$$

where $\mathcal{L}_{\text{edit}} = \text{mean}(|\hat{z} - z_t|)$. Complete definitions of the regional objectives are provided in Appendix A.1.

Once trained, the Privacy Masker is independently applied to video frames. Because its predictions operate over spatial visual features rather than directly at the pixel level, we treat them as coarse semantic proposals for subsequent spatial and temporal refinement.

3.2 Mask Selection and Temporal Refinement

Frame-level predictions provide semantic localization of privacy-sensitive content but can have coarse boundaries and vary across neighboring frames. Moreover, sensitive content may appear differently across shots. We therefore perform shot-aware seed selection followed by spatial refinement and temporal propagation.

We first partition each video into shots using OmniShotCut (Wang et al., 2026). Within each shot, the Privacy Masker predictions are used to select a representative seed mask. SAM2 (Ravi et al., 2025) then refines the coarse region and propagates it throughout the shot, yielding spatially precise and temporally consistent masks. The Privacy Masker therefore determines *what* should be treated as privacy-sensitive, while SAM2 determines its precise spatial extent and tracks it over time. Seed selection and refinement details are provided in Appendix A.2.

3.3 Privacy-Preserving Video Transformation

The resulting privacy masks support two forms of video transformation. **Mask-based anonymization** directly hides the identified regions, providing a suppression-based transformation in which sensitive visual evidence is removed. **Generative anonymization** instead replaces the identified content with plausible alternatives while preserving the surrounding visual information.

For generative anonymization, we use EditCtrl (Litman et al., 2026), providing the temporally consistent masks as editable regions together with a simple action-conditioned prompt. Importantly, the video editor itself does not determine what is privacy-sensitive; it operates only on regions localized by Mask2Safe. This allows us to separately examine the effect of *where* privacy transformations are applied and whether replacing the localized content better preserves meaningful visual information than directly suppressing it. Generation settings and prompts are provided in Appendix A.2.

4 SAFEQA: EVALUATING MEANINGFUL ANONYMIZATION

Existing privacy-preserving video methods are commonly evaluated through the trade-off between privacy-attribute prediction and downstream utility (Dave et al., 2022; Li et al., 2023; Fioresi et al., 2023; Aslam et al., 2026). While these metrics measure whether sensitive attributes remain predictable and whether the overall activity remains recognizable, they do not directly assess whether *specific non-sensitive visual information* remains observable after anonymization. A method may therefore achieve low privacy leakage by broadly masking or degrading visual content while retaining sufficient global cues for action recognition.

To address this limitation, we introduce **SafeQA**, a video question-answering benchmark for evaluating what information remains recoverable after anonymization. Given an original video, SafeQA constructs privacy-grounded question triplets and establishes their ground-truth answers from the original video. The same questions are then answered using its anonymized versions. As illustrated in Fig. 1(b), each triplet evaluates three complementary dimensions: *Privacy*, whether the original sensitive information remains recoverable; *Visibility*, whether the associated visual content remains observable; and *Local Utility*, whether nearby non-sensitive information remains recoverable. By reporting these dimensions separately, SafeQA distinguishes privacy protection from preservation and reveals whether privacy is achieved by selectively transforming sensitive information or by broadly suppressing useful visual content.

4.1 SAFEQA CONSTRUCTION

Given an original video, its downstream action label, and privacy categories derived from VISPR (Orekondy et al., 2017), we use Google Gemini 3.8 Flash (Gemini Team, 2023) to construct privacy-grounded question triplets from visual evidence in the video. The model first identifies clearly observable privacy-sensitive information whose modification would not alter the essential semantics of the video. For each valid privacy question (Q1), it constructs a corresponding visibility question (Q2) and local utility question (Q3) grounded in the associated visual content.

Question generation is designed to be conservative: candidate privacy information must be clearly observable in the original video, visually grounded, and distinguishable from information necessary to recognize the downstream action. Videos without suitable privacy-sensitive information need not produce a triplet, while privacy-rich videos may contribute multiple non-redundant triplets. We retain up to three triplets per video.

All questions and ground-truth answers are established from the original video and used without modification to evaluate its anonymized versions. Question generation relies only on visual evidence, without audio, dialogue, subtitles, or external knowledge. SafeQA contains $3,005$ question triplets across $2,861$ videos. Please see Appendix C.1 for the complete generation criteria and prompts, privacy categories, and additional benchmark statistics.

4.2 QUESTION TYPES

Q1: Privacy (↓). Privacy questions evaluate whether specific privacy-sensitive information present in the original video remains recoverable after anonymization. The answer must depend on visual evidence from the specific video rather than contextual priors or information required to recognize the downstream action. We additionally exclude incidental visual properties whose modification would not meaningfully alter the underlying privacy-sensitive information. The ground-truth answer is established from the original video and retained when evaluating its anonymized version. Any prediction that differs from the original ground-truth answer is treated as non-recovery; thus, the metric evaluates whether the *original* sensitive information remains recoverable rather than whether anonymization produces a particular replacement value.

Q2: Visibility (↑). Privacy protection alone does not indicate whether the visual content associated with the sensitive information remains observable. Visibility questions are binary *Yes/No* questions that assess whether the relevant person, object, or region remains visually accessible after anonymization. For example, if a privacy question concerns an identifying feature on a person's arm, the corresponding visibility question may ask whether that arm remains visible.

Q3: Local Utility (↑). Visibility alone does not guarantee that useful visual information is preserved: the affected content may remain visible while nearby semantic information about an action or interaction is lost. Local utility questions evaluate whether specific *non-sensitive* information associated with the privacy-sensitive content remains recoverable after anonymization, such as what a person is doing with an affected body part or which object they are interacting with. Unlike conventional action recognition, which evaluates whether the overall activity remains identifiable, local utility measures the preservation of fine-grained information associated with the anonymized content and its surrounding context.

Please see Appendix C.2 for detailed question-generation criteria and additional examples.

4.3 Evaluation

Given an anonymized video, an answering VLM answers the same questions whose ground-truth answers were established from the original video. We report accuracy separately for *Privacy* (↓), *Visibility* (↑), and *Local Utility* (↑). Reporting the three dimensions separately reveals how a transformation succeeds or fails: low Privacy accuracy may result either from selectively changing the sensitive information or from suppressing the associated visual content altogether, while Visibility and Local Utility distinguish whether useful visual information remains accessible. We additionally evaluate global video utility using conventional action-recognition accuracy in Sec. 5. Please see Appendix C.3 for the complete evaluation protocol and answering prompts.

5 Experiments

We evaluate Mask2Safe from three complementary perspectives. First, we examine the conventional privacy–utility trade-off by measuring privacy leakage and downstream action recognition on anonymized videos. Second, we evaluate whether action classifiers trained on anonymized videos transfer to the original video domain, providing a complementary measure of the task-relevant distribution shift introduced by anonymization. Finally, we use SafeQA to examine whether anonymization prevents recovery of the original privacy-sensitive information while preserving the associated visual content and specific non-sensitive information.

5.1 Experimental Setup

Datasets and Baselines. We evaluate on VP-UCF101 and VP-HMDB51 (Li et al., 2023), derived from UCF101 (Soomro et al., 2012) and HMDB51 (Kuehne et al., 2011), respectively. We compare against spatial downsampling at $2\times$ and $4\times$, TeD-SPAD (Fioresi et al., 2023), and Unsafe2Safe (Dinh & Jin, 2026). For a controlled comparison, we generate anonymized videos using each method and train all downstream action and privacy classifiers using the same architecture, data splits, and training protocol. We include prior methods for which the released resources permit reproduction of the anonymized videos required by our evaluation. We evaluate two variants of Mask2Safe: mask-based anonymization and generative anonymization.

Evaluation Metrics. We evaluate conventional privacy preservation using privacy cMAP and F1 (↓) and downstream utility using action-recognition Top-1 accuracy (↑). We additionally measure *anonymized-to-original action transfer*: an action classifier trained on anonymized training videos is evaluated on both anonymized and original test videos. We report the absolute difference $|\Delta|$ between the two Top-1 accuracies, where a smaller gap indicates less task-relevant domain shift. We evaluate privacy prediction on anonymized videos, where it directly measures the sensitive information remaining after anonymization.

For SafeQA, we report Privacy (↓), Visibility (↑), and Local Utility (↑) accuracy. See Sec. 4 for the complete benchmark definition and evaluation protocol.

Implementation Details. The Privacy Masker uses a frozen DINOv2 ViT-S/14 (Oquab et al., 2024) encoder and a lightweight trainable mask decoder. We use SAM2 for temporal mask refinement and EditCtrl for generative anonymization. Complete training details, model configurations, anonymization settings, and action/privacy classifier architectures are provided in Appendix B.

Table 1: **Privacy–utility and anonymized-to-original transfer.** We report action-recognition Top-1 accuracy (↑) on anonymized videos and, for classifiers trained on anonymized videos, their accuracy on the original test videos. $|\Delta|$ is the absolute difference between the two (↓). Privacy leakage is measured by cMAP and F1 (↓). **Bold** and underlined values indicate the best and second-best results, respectively, among anonymization methods.

		Action Utility			Privacy	
Dataset	**Method**	**Anon. ↑**	**Original ↑**	**$\|\Delta\|$ ↓**	**cMAP ↓**	**F1 ↓**
VP-UCF101	Raw	81.62	–	–	72.86	0.695
	Downsample 2×	79.38	77.13	2.25	72.53	0.686
	Downsample 4×	74.91	73.54	1.37	71.61	0.679
	TeD-SPAD	62.58	50.09	12.49	**69.85**	**0.662**
	Unsafe2Safe	**85.14**	65.33	19.81	72.11	0.693
	Mask2Safe (Mask-based)	67.15	67.02	**0.13**	70.04	0.664
	Mask2Safe (Generative)	74.53	**76.51**	1.98	71.07	0.681
VP-HMDB51	Raw	48.04	–	–	72.78	0.684
	Downsample 2×	41.44	43.92	2.48	70.81	0.683
	Downsample 4×	38.89	39.28	0.39	71.45	0.678
	TeD-SPAD	30.65	22.55	8.10	67.85	**0.633**
	Unsafe2Safe	**48.50**	37.39	11.11	69.31	0.661
	Mask2Safe (Mask-based)	35.49	34.05	1.44	**67.59**	0.648
	Mask2Safe (Generative)	39.93	**39.08**	**0.85**	68.70	0.650

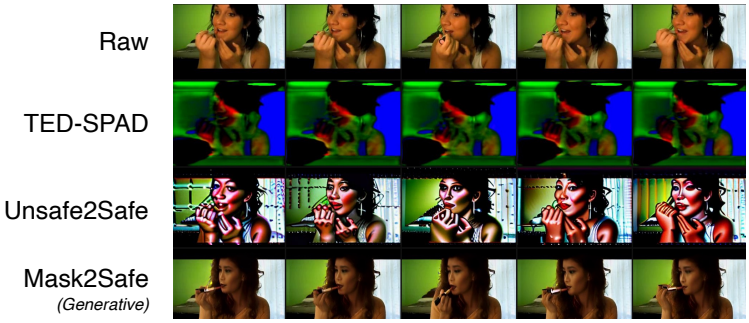

Figure 4: **Qualitative comparison on VP-UCF101.** Frames from *ApplyLipstick*. Compared with prior methods, Mask2Safe localizes its modifications while better preserving surrounding visual content across frames. Additional examples and intermediate masks are provided in Appendix D.

5.2 Results

Privacy–Utility and Action Transfer. Table 1 reveals that conventional privacy and utility metrics capture only part of the behavior of privacy-preserving transformations. Stronger suppression of privacy information does not necessarily coincide with better preservation of action utility, while high action accuracy alone does not imply strong privacy protection. Our two variants expose this trade-off directly: mask-based anonymization more aggressively suppresses visual information, whereas generative anonymization better retains action-relevant content by replacing rather than removing the localized regions.

The transfer evaluation reveals a complementary distinction that is not captured by in-domain action accuracy. Existing anonymization methods can retain strong action recognition when classifiers are trained and tested on similarly transformed videos, yet transfer substantially worse to the original videos. This suggests that successful in-domain recognition can partly rely on characteristics specific to the anonymized domain. In contrast, classifiers trained on Mask2Safe-anonymized videos exhibit substantially smaller changes when evaluated on the original domain. In particular, generative anonymization combines strong action utility with a small transfer gap, suggesting that localized

Table 2: **Evaluation on SafeQA.** Privacy (↓) measures recovery of the original sensitive information; Visibility (↑) measures whether the affected visual content remains observable; and Local Utility (↑) measures preservation of nearby non-sensitive information.

Dataset	Method	Privacy ↓	Visibility ↑	Local Utility ↑
VP-UCF101	Raw	97.29	98.07	94.26
	Downsample 2×	91.18	97.29	91.37
	Downsample 4×	78.42	95.59	87.42
	TeD-SPAD	55.00	68.78	66.44
	Unsafe2Safe	**50.55**	71.72	65.93
	Mask2Safe (Mask-based)	62.30	50.64	**74.93**
	Mask2Safe (Generative)	63.77	**77.13**	73.60
VP-HMDB51	Raw	97.94	98.55	94.56
	Downsample 2×	84.16	86.70	85.49
	Downsample 4×	74.18	86.79	82.06
	TeD-SPAD	58.40	65.54	68.92
	Unsafe2Safe	**49.09**	71.95	64.57
	Mask2Safe (Mask-based)	62.03	47.16	73.28
	Mask2Safe (Generative)	70.98	**80.29**	**77.27**

replacement can preserve action-relevant characteristics without introducing a strongly distinct task-relevant domain. Figure 4 provides a visual comparison of the anonymized outputs.

Does Privacy Preservation Retain Meaningful Visual Information? Table 2 further demonstrates why reduced privacy predictability alone is insufficient for evaluating meaningful anonymization. A transformation can prevent recovery of sensitive information by making the associated visual content difficult to observe, rather than by selectively changing the sensitive information itself. SafeQA distinguishes these behaviors by separately examining Privacy, Visibility, and Local Utility.

The comparison between mask-based and generative anonymization is particularly informative because both operate on privacy-sensitive regions localized by the same framework and differ primarily in how those regions are transformed. Mask-based anonymization achieves strong Local Utility despite substantially lower Visibility, indicating that nearby non-sensitive information can remain recoverable even when the localized regions themselves are suppressed. Generative anonymization substantially improves Visibility while maintaining strong Local Utility, achieving the highest Visibility among anonymization methods on both datasets and the highest Local Utility on VP-HMDB51. These results show that replacing localized sensitive content can retain observable visual content and nearby useful information more effectively than simply hiding the affected regions.

6 Conclusion

We presented **Mask2Safe**, a privacy-preserving video generation framework that transfers semantic privacy localization learned from unsafe–safe image pairs to video. A lightweight **Privacy Masker** identifies privacy-sensitive regions without frame-wise VLM reasoning, and shot-aware SAM2 refinement turns its predictions into temporally consistent masks supporting mask-based or generative anonymization. We further introduced **SafeQA**, a benchmark that evaluates anonymization at the content level by measuring recovery of sensitive information, visibility of affected content, and preservation of nearby non-sensitive information. On VP-UCF101 and VP-HMDB51, Mask2Safe reduces privacy leakage while retaining action utility and exhibits substantially smaller anonymized-to-original transfer gaps than prior anonymization methods. SafeQA shows that generative anonymization substantially improves visibility over mask-based anonymization while maintaining strong local utility, highlighting the benefit of replacing rather than suppressing localized sensitive content. These results expose a remaining trade-off between removing sensitive information and preserving useful visual content, leaving meaningful anonymization an open challenge. SafeQA primarily focuses on content-level preservation and does not explicitly evaluate temporal consistency as a separate dimension, motivating future evaluation of privacy-preserving transformations requiring temporal evidence.

AI USE STATEMENT

The authors used generative AI tools to assist with language editing and improving the clarity and readability of the manuscript. All technical content, research ideas, experimental design, implementation, analysis, and conclusions were developed and verified by the authors. The authors take full responsibility for the content of the paper.

Generative AI and vision-language models used as components of the proposed method, benchmark construction, or evaluation pipeline are part of the research methodology and are described explicitly in the paper.

ETHICS STATEMENT

This work studies privacy-preserving video generation with the goal of reducing the exposure of sensitive visual information while preserving information useful for video understanding. By design, Mask2Safe localizes privacy-sensitive regions and supports either mask-based or generative anonymization of the affected content. Nevertheless, privacy-preserving generation introduces several ethical considerations.

First, no automated anonymization method can guarantee complete removal of sensitive information. Sensitive attributes may remain recoverable from unmodified regions, temporal context, audio, correlations across frames, or other contextual cues not explicitly addressed by the system. The privacy evaluations reported in this work should therefore be interpreted as empirical measures of privacy protection rather than guarantees of anonymity.

Second, generative replacement may introduce visual content that was not present in the original video. Although the objective is to preserve non-sensitive semantics and downstream utility, generated content should not be interpreted as a faithful record of the original sensitive regions. This is particularly important in applications where visual authenticity or evidentiary integrity is required.

Third, definitions of sensitive information are context dependent and may vary across individuals, cultures, and applications. SafeQA and the privacy attributes considered in our experiments provide an evaluation framework rather than an exhaustive definition of privacy.

Our experiments use established research datasets and are conducted for the purpose of developing and evaluating privacy-preserving video representations and generation. We do not attempt to identify individuals depicted in the datasets. We encourage future applications of this technology to consider dataset consent, appropriate access controls, domain-specific definitions of sensitive information, and the limitations of automated anonymization.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the Mask2Safe framework, including the Privacy Masker, temporal mask propagation and refinement, and mask-based and generative anonymization pipelines. The paper also describes the construction and evaluation protocol of SafeQA, including its Privacy, Visibility, and Local Utility criteria. Experimental settings, datasets, evaluation metrics, implementation details, and comparisons with prior privacy-preserving video methods are documented in the main paper and supplementary material.

Experiments are conducted on the publicly available VP-UCF101 and VP-HMDB51 benchmarks. We report results using established privacy and action-recognition metrics together with the proposed SafeQA evaluation to enable comparison across complementary dimensions of privacy and utility.

Code, evaluation scripts, SafeQA annotations, prompts, and configuration files necessary to reproduce the reported experiments will be released upon publication, subject to the licenses and distribution requirements of the underlying datasets and pretrained models. For third-party models or resources that cannot be redistributed, we will provide instructions and references for obtaining the corresponding resources from their original sources.

REFERENCES

Nazia Aslam and Kamal Nasrollahi. Balancing privacy and action performance: A penalty-driven approach to image anonymization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 729–738, June 2025.

Nazia Aslam, Abhisek Ray, Joakim Bruslund Haurum, Lukas Esterle, and Kamal Nasrollahi. From pixels to privacy: Temporally consistent video anonymization via token pruning for privacy preserving action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 8580–8589, June 2026.

Asfandyar Azhar, Nidhish Shah, Shaurjya Mandal, and Yongjie Jessica Zhang. Privacy preservation using superimposed 3d-models for self-supervised training in action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 2304–2314, October 2025.

Adith Boloor, Weikai Lin, Tianrui Ma, Yu Feng, Yuhao Zhu, and Xuan Zhang. Privateeye: In-sensor privacy preservation through optical feature separation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2357–2367. IEEE, 2025.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, 2023.

Lixin Chen, Chaomeng Chen, Jiale Zhou, Zhijian Wu, and Xun Lin. Stegavar: Privacy-preserving video action recognition via steganographic domain analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2930–2938, 2026.

Ishan Rajendrakumar Dave, Chen Chen, and Mubarak Shah. Spact: Self-supervised privacy preservation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20164–20173, June 2022.

Minh Dinh and SouYoung Jin. Unsafe2safe: Controllable image anonymization for downstream utility. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3326–3336, June 2026.

Anil Egin, Andrea Tangherloni, and Antitza Dantcheva. Now you see me, now you don't: A unified framework for expression consistent anonymization in talking head videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 5984–5993. IEEE, 2025.

Joseph Fioresi, Ishan Rajendrakumar Dave, and Mubarak Shah. Ted-spad: Temporal distinctiveness for self-supervised privacy-preservation for video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13598–13609, October 2023.

Joseph Fioresi, Ishan Rajendrakumar Dave, and Mubarak Shah. Privacy beyond pixels: Latent anonymization for privacy-preserving video understanding. In *International Conference on Learning Representations (ICLR)*, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Google Gemini Team. Gemini: A family of highly capable multimodal models, 2023. URL `https://arxiv.org/abs/2312.11805`.

Carlos Hinojosa, Miguel Marquez, Henry Arguello, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Privhar: Recognizing human actions from privacy-preserving lens. In *European Conference on Computer Vision (ECCV)*, pp. 314–332. Springer, 2022.

Filip Ilic, He Zhao, Thomas Pock, and Richard P Wildes. Selective, interpretable and motion consistent privacy attribute obfuscation for action recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18730–18739. IEEE, 2024.

Yuchi Ishikawa, Masayoshi Kondo, and Hirokatsu Kataoka. Learnable cube-based video encryption for privacy-preserving action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7003–7013, January 2024.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202, 2025.

H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *International Conference on Computer Vision (ICCV)*, pp. 2556–2563, 2011. doi: 10.1109/ICCV.2011.6126543.

Sudhakar Kumawat and Hajime Nagahara. Privacy-preserving action recognition via motion difference quantization. In *European Conference on Computer Vision (ECCV)*, pp. 518–534. Springer, 2022.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22195–22206, June 2024.

Ming Li, Xiangyu Xu, Hehe Fan, Pan Zhou, Jun Liu, Jia-Wei Liu, Jiahe Li, Jussi Keppo, Mike Zheng Shou, and Shuicheng Yan. Stprivacy: Spatio-temporal privacy-preserving action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5106–5115, October 2023.

Yehonathan Litman, Shikun Liu, Dario Seyb, Nicholas Milef, Yang Zhou, Carl Marshall, Shubham Tulsiani, and Caleb Leak. Editctrl: Disentangled local and global control for real-time generative video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8965–8975, June 2026.

Maxim Maximov, Ismail Elezi, and Laura Leal-Taixe. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Saemi Moon, Myeonghyeon Kim, Zhenyue Qin, Yang Liu, and Dongwoo Kim. Anonymization for skeleton action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 15028–15036, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=Ha6RTeWMd0.

Zhongzheng Ren, Yong Jae Lee, and Michael S. Ryoo. Learning to anonymize faces for privacy preserving action detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

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

Michael S. Ryoo, Brandon Rothrock, Charles Fleming, and Hyun Jong Yang. Privacy-preserving human activity recognition from extreme low resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pp. 4255–4262, 2017. doi: 10.1609/aaai.v31i1.11233.

David Schneider, Sina Sajadmanesh, Vikash Sehwag, Saquib Sarfraz, Rainer Stiefelhagen, Lingjuan Lyu, and Vivek Sharma. Activity recognition on avatar-anonymized datasets with masked differential privacy. *arXiv preprint arXiv:2410.17098*, 2024.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *CRCV-TR-12-01*, 2012.

Adam Sun, Eshaan Barkataki, Arnold Milstein, Gordon Wetzstein, and Ehsan Adeli. Regenhuman: Re-generating human appearances for realistic full-body video anonymization. *arXiv preprint arXiv:2606.14972*, 2026.

Daksh Thapar, Aditya Nigam, and Chetan Arora. Anonymizing egocentric videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2320–2329, October 2021.

Matteo Tomei, Lorenzo Baraldi, Simone Bronzin, and Rita Cucchiara. Estimating (and fixing) the effect of face obfuscation in video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 3263–3269, June 2021.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning (ICML)*, pp. 10347–10357. PMLR, 2021.

Boyang Wang, Guangyi Xu, Jiahui Zhang, Zhipeng Tang, and Zezhou Cheng. Omnishotcut: Holistic relational shot boundary detection with shot-query transformer. *arXiv preprint arXiv:2604.24762*, 2026.

Zihao W. Wang, Vibhav Vineet, Francesco Pittaluga, Sudipta N. Sinha, Oliver Cossairt, and Sing Bing Kang. Privacy-preserving action recognition using coded aperture videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Zhenyu Wu, Haotao Wang, Zhaowen Wang, Hailin Jin, and Zhangyang Wang. Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2126–2139, 2022. doi: 10.1109/TPAMI.2020.3026709.

Zhi-Wei Xia, Kun-Yu Lin, Yuan-Ming Li, Wei-Jin Huang, Xian-Tuo Tan, and Wei-Shi Zheng. Less static, more private: Towards transferable privacy-preserving action recognition by generative decoupled learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12894–12903, October 2025.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9772–9781. IEEE, 2021.

Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, Jianke Zhu, and Lidong Bing. Videorefer suite: Advancing spatial-temporal object understanding with video llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18970–18980, June 2025.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 31428–31449. Curran Associates, Inc., 2023. doi: 10.52202/075280-1365. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/64008fa30cba9b4d1ab1bd3bd3d57d61-Paper-Datasets_and_Benchmarks.pdf.

APPENDIX

A MASK2SAFE DETAILS

A.1 PRIVACY MASKER TRAINING

We provide additional details for the Privacy Masker introduced in Sec. 3.1. Given an unsafe–safe image pair (I_s, I_t), we encode both images using the frozen VAE encoder $\mathcal{E}$ to obtain z_s and z_t. The Privacy Masker predicts a soft spatial mask m from frozen DINOv2 source features, and a replacement network predicts the content used to reconstruct the privacy-preserving target.

Replacement Prediction. Given the source latent z_s and predicted soft mask m, a text-free U-Net G produces a candidate replacement latent. To prevent the replacement pathway from directly optimizing the Privacy Masker through its mask input, we detach the mask before providing it to G:

$$r = G\big([z_s; \mathrm{sg}(m)]\big), \tag{5}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation. The predicted mask then composites the replacement with the source:

$$\hat{z} = m \odot r + (1 - m) \odot z_s. \tag{6}$$

Because m is not detached in this composition, reconstruction gradients provide supervision to the Privacy Masker by encouraging it to expose regions whose modification is necessary to reproduce the target.

Reconstruction Objective. We supervise the composited latent using the mean absolute error (MAE) to the target latent:

$$\mathcal{L}_{\mathrm{edit}} = \mathrm{mean}\left(|\hat{z} - z_t|\right) = \frac{1}{|\hat{z}|}\|\hat{z} - z_t\|_1. \tag{7}$$

For edited source–target pairs, this objective encourages the mask to expose regions that must change to reconstruct the privacy-preserving target. For pairs requiring no privacy modification, where $I_s = I_t$, it instead encourages the mask to remain closed and preserve the source content.

Regional Objectives. The reconstruction objective alone does not explicitly constrain how the replacement latent behaves inside and outside the predicted region. For a spatial weighting map w and element-wise distance A, we define the weighted reduction

$$\mathcal{M}_w(A) = \frac{\sum(w \odot A)}{C \sum w + \epsilon}, \tag{8}$$

where C is the number of latent channels and $\epsilon = 10^{-6}$.

Outside the predicted mask, the replacement should remain consistent with the source:

$$\mathcal{L}_{\mathrm{preserve}} = \mathcal{M}_{1-\mathrm{sg}(m)}\left(|r - z_s|\right). \tag{9}$$

Within the predicted mask, we encourage the replacement to be closer to the privacy-preserving target than to the original source:

$$\mathcal{L}_{\mathrm{pref}} = \mathcal{M}_{\mathrm{sg}(m)}\left(\left[|r - z_t| - |r - z_s|\right]_+\right), \tag{10}$$

where $[\cdot]_+ = \max(\cdot, 0)$. The mask weights are detached in both regional objectives so that these terms optimize the replacement pathway without directly changing the Privacy Masker.

Mask Regularization. Without an explicit size constraint, the Privacy Masker can explain the source–target transformation by exposing unnecessarily large regions. We therefore regularize the spatial extent of the predicted mask using

$$\mathcal{L}_{\mathrm{area}} = \mathrm{mean}(m). \tag{11}$$

The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{edit}} + \mathcal{L}_{\mathrm{preserve}} + \mathcal{L}_{\mathrm{pref}} + \lambda_{\mathrm{area}}\mathcal{L}_{\mathrm{area}}, \tag{12}$$

with $\lambda_{\mathrm{area}} = 0.05$ in all experiments.

A.2 Implementation Details

Privacy Masker Architecture. A frozen DINOv2 ViT-S/14 (Oquab et al., 2024) extracts a $384 \times 16 \times 16$ feature grid from the source image resized to 224×224. A lightweight decoder upsamples this grid with a stride-2 transposed convolution with 16 output channels, followed by GroupNorm, SiLU, and a 1×1 convolution, producing a single-channel 32×32 mask logit. We then apply a sigmoid activation so that the mask values lie in the range $[0, 1]$:

$$m = \sigma(f_\theta(F_s)) \,. \tag{13}$$

The replacement network G is a U-Net initialized from the MagicBrush (Zhang et al., 2023) EMA weights with its text cross-attention modules removed; its input is the source latent concatenated with the detached mask. The Stable Diffusion (Rombach et al., 2022) VAE and DINOv2 remain frozen, while the mask decoder and G are optimized jointly.

Privacy Masker Training. We train on the Unsafe2Safe dataset (Dinh & Jin, 2026), built on MSCOCO, using both its edited (unsafe→safe) pairs and its unedited images, which form identity pairs with $I_t = I_s$. Following InstructPix2Pix (Brooks et al., 2023), the source and target undergo identical augmentation: resizing to 286×286, random 256×256 cropping, and horizontal flipping with probability 0.5. The frozen VAE maps each image to a $4 \times 32 \times 32$ latent. We use AdamW with a learning rate of 10^{-4} and a cosine schedule defined over 10,000 steps with a final rate of 10^{-5}, bfloat16 precision, and a batch size of 96 per GPU on eight GPUs. Following MagicBrush, we train for 5,000 optimizer steps and use the final checkpoint.

Frame-Level Mask Inference. At inference time, the Privacy Masker is applied independently to every video frame, resized to 256×256. The predicted 32×32 soft mask is bilinearly upsampled to the original frame resolution.

Shot-Aware Seed Selection and Temporal Refinement. We partition each video into shots using OmniShotCut (Wang et al., 2026) in its default mode, merging intervals labeled as transitions or padding into an adjacent shot. For each shot S_k, we threshold the coarse masks at $\tau = 0.5$ and select the frame with the largest positive mask area:

$$t_k^\star = \arg\max_{t \in S_k} \sum_{x,y} \mathbf{1}\big[m^t(x, y) \geq \tau\big] \,. \tag{14}$$

Shots without positive mask pixels are skipped. The selected binary masks are added as mask prompts to a single SAM2.1 Hiera-Large (Ravi et al., 2025) session per video, with one object per shot. We propagate all objects forward and backward over the entire video, keep each object's predictions (positive logits) only within its own shot, and take the union over objects to obtain a binary privacy mask for every frame.

Mask-Based Anonymization. Pixels inside the refined masks are set to black at the original video resolution. Videos without any positive mask are left unchanged.

Generative Anonymization. We use EditCtrl (Litman et al., 2026) with its released local and global control weights on the Wan2.1-VACE-1.3B backbone (Jiang et al., 2025), enabling both local and global inpainting. Frames and masks are center-cropped to a 4:3 aspect ratio and resized to 320×240. Masks are binarized, and masked pixels in the conditioning frames are set to black. Each video is edited in a single pass, padded to the required $4k+1$ frames by repeating its last frame without a mask, using 25 denoising steps, a classifier-free guidance scale of 5, and a denoising strength of 1.0. The text prompt is `A video of {class_name}.`, where `class_name` is obtained by splitting the action-class directory name into words (e.g., `ApplyEyeMakeup` → "Apply Eye Makeup"). We use the negative prompt `long body, low-res, bad anatomy, bad hands, missing fingers, extra digit, fewer digits, cropped, worst quality, low quality, deformed body, bloated, ugly, blurry, low res, unaesthetic`. Edited videos are saved at 320×240 with the original frame rate; videos without any positive mask are passed through unchanged.

B EXPERIMENTAL DETAILS

This section describes the experimental setup; the implementation of Mask2Safe itself is detailed in Appendix A.2. Experiments are conducted using NVIDIA A100-SXM4-80GB and RTX 6000 Ada Generation 48GB GPUs.

B.1 BASELINES

Spatial downsampling. Each frame is spatially downsampled by a factor of 2 or 4 (e.g., from 320×240 to 160×120 or 80×60 for UCF101) and stored at the reduced resolution. The downstream classifiers resize the frames back to their input resolution.

TeD-SPAD. TeD-SPAD (Fioresi et al., 2023) extends the self-supervised anonymization framework of SPAct (Dave et al., 2022) to privacy-preserving video anomaly detection. Specifcally, its UNet++ anonymizer is trained adversarially against a self-supervised privacy branch on VISPR (Orekondy et al., 2017) images, while an action-recognition utility branch on UCF101 is combined with a temporally distinct triplet loss that preserves the temporal features required for anomaly detection. We use the anonymizer checkpoint released with the official code, without retraining, for both VP-UCF101 and VP-HMDB51. The anonymizer is applied frame by frame at 224×224, and its output is resized to the original frame resolution. Since TeD-SPAD was designed for anomaly detection, this baseline also examines how an existing learned anonymizer transfers to a new privacy–utility evaluation setting.

Unsafe2Safe. Since Unsafe2Safe (Dinh & Jin, 2026) was designed for images, we adapt it to video while at the same time keeping its captions consistent over the video. For each video, Gemini 3.5 Flash-Lite (Gemini Team, 2023), working as the VLM in Unsafe2Safe pipeline, receives the entire video and produces a privacy-safe public caption of its content, from which the model then derives the corresponding privacy-preserving edit instruction. Editing is performed frame by frame with the released SafeAttention editing model of Unsafe2Safe, which was built on InstructPix2Pix (Brooks et al., 2023). Following the reported implementation details, every frames are resized so that their longest side is 256 pixels and edited with 25 sampling steps with text and image guidance scales of 7.5 and 1.5. Once edited, the output frames are resized back to the original resolution.

B.2 ACTION AND PRIVACY EVALUATION PROTOCOL

Following STPrivacy (Li et al., 2023), we train a separate action classifier and privacy classifier on the training videos of each representation being compared. Both use a ViT-S/16 video transformer initialized from ImageNet-pretrained DeiT-S (Touvron et al., 2021) weights, with the 2D patch embedding inflated to $2 \times 16 \times 16$ tubelets by averaging, and all parameters are fine-tuned. Each training clip has 16 frames sampled with a stride of four for UCF101 and two for HMDB51, resized to 112×112. We train each classifier for 80 epochs using AdamW with a weight decay of 0.05, a batch size of 32, a learning rate of $10^{-3} \times \text{batch size}/512$ with cosine annealing, and mixed precision. The action classifier uses a softmax cross-entropy head, and the privacy classifier predicts the five annotated privacy attributes with sigmoid outputs and a binary cross-entropy loss.

At test time, predictions are averaged over five temporal clips and three spatial crops with a crop scale of 0.8, and we report the final checkpoint after 80 epochs. We report Top-1 accuracy for action recognition and class-averaged F1 and mean average precision (cMAP) for the five privacy attributes. Higher action accuracy indicates better utility, whereas lower privacy-recognition scores indicate less recoverable private information.

Anonymized-to-original transfer. For the transfer evaluation in Table 1, the action classifier trained on each method's anonymized training videos is evaluated, without any adaptation, on the original test videos using the same protocol. $|\Delta|$ is the absolute difference between its Top-1 accuracy on the anonymized and on the original test videos.

VLM for SafeQA. SafeQA questions are generated from the original videos with Gemini 3.8 Flash, as described in Appendix C.1, and all questions are answered with Gemini 3.5 Flash-Lite. The

- **People**: faces (full or partial), visible skin in private contexts, tattoos, or emotional expressions
- **Appearance cues**: clothing, accessories, or physical features revealing or distinguishing gender, culture, religion, status, identity, affiliation, or role; distinctive clothing patterns, styles, or hairstyles
- **Health indicators**: casts, wheelchairs, bandages, medications, disabilities, or visible illness/injury
- **Texts or symbols**: names, receipts, documents, screens, handwriting, license plates, usernames, phone numbers, or identifying logos, symbols, or affiliation information
- **Locations or timestamps**: homes, offices, bedrooms, classrooms, addresses, GPS data, clocks, dates, or timestamps; cues revealing a private or identifying location
- **Personal interactions**: families, friends, small groups, teams, close relationships, or interactions revealing personal associations
- **Professional or school settings**: uniforms, tools, name tags, diplomas, workspaces, or institutional/school affiliation cues
- **Cultural, political, or religious identifiers**: flags, slogans, attire, ceremonies, posters, or related symbols and affiliations
- **Private belongings or behaviors**: personal photos, mail, social media, travel documents, receipts, hobbies, or objects revealing a person's private life

Figure 5: Privacy-related visual criteria used for SafeQA question generation, following the VISPR-derived criteria of Unsafe2Safe (Dinh & Jin, 2026).

questions and their ground-truth answers are fixed from the original videos and remain unchanged across all anonymization methods.

C SafeQA Details

This appendix provides additional details on the construction and evaluation of SafeQA, complementing Sec. 4 of the main paper.

C.1 Benchmark Construction

Question Generation. Given an original video, its downstream action label, and a predefined set of privacy categories derived from VISPR (Orekondy et al., 2017), we use Google Gemini (Gemini Team, 2023) 3.8 Flash to construct privacy-grounded question triplets from the visual content of the video. Each triplet consists of a privacy question (Q1), a visibility question (Q2), and a local utility question (Q3).

Question generation is designed to be conservative. The model first examines the video for clearly observable privacy-sensitive information across the predefined privacy categories. A candidate is retained only when its value can be determined reliably from visual evidence in the video, is not simply implied by the downstream action or contextual priors, and can be modified without changing the essential semantics of the video. Incidental visual properties, such as generic colors or unrelated visual details, are not considered sufficient privacy information by themselves.

For each valid privacy candidate, the model generates a multiple-choice privacy question (Q1) together with its ground-truth answer and supporting visual evidence. It then constructs a corresponding visibility question (Q2) that assesses whether the associated person, object, or region remains observable, and a local utility question (Q3) that evaluates whether related non-sensitive visual information remains recoverable.

A video is not required to produce a question triplet. If no privacy-sensitive information satisfies the generation criteria, no triplet is retained for that video. Conversely, privacy-rich videos may contribute multiple triplets when they capture distinct, non-redundant privacy-sensitive information. We retain up to three triplets per video.

All questions and ground-truth answers are established from the original video and subsequently used without modification when evaluating its anonymized versions. Question generation relies only on visual evidence contained in the video and does not use audio, dialogue, subtitles, or external knowledge.

Benchmark Statistics. Starting from $5,313$ candidate videos from VP-UCF101 and VP-HMDB51, our generation procedure produces $3,005$ question triplets across $2,861$ videos. Table 3 provides the dataset-level statistics.

Table 3: **SafeQA benchmark statistics.** We report the number of candidate videos, videos contributing at least one valid triplet, and resulting question triplets.

Dataset	Candidate Videos	Valid Videos	Question Triplets
VP-UCF101	3,783	2,086	2,178
VP-HMDB51	1,530	775	827
Total	5,313	2,861	3,005

You are constructing a high-quality benchmark for evaluating privacy-preserving video anonymization.
You are given: (1) an ORIGINAL, UNANONYMIZED video, and (2) its action label: {`label`}
Your task is to determine whether this video supports ZERO OR MORE high-confidence privacy-grounded QA triplets. Each triplet consists of Q1: Privacy, Q2: Visibility, and Q3: Local Utility. Generate AT MOST 3 triplets per video. **QUALITY IS MORE IMPORTANT THAN COVERAGE.** Zero, one, two, or three triplets are all valid outcomes. Do NOT generate a triplet merely because a video was provided. Do NOT generate additional triplets merely to reach three. When evidence is weak, ambiguous, too small, poorly visible, semantically essential, or only marginally privacy-relevant, reject that candidate. All questions and ground-truth answers must be established solely from DIRECTLY VISIBLE evidence in the ORIGINAL video.

—

PRIVACY TAXONOMY
Search for privacy-relevant information from ALL of the following categories:
`[PRIVACY CRITERIA]`
The taxonomy defines WHAT TYPES OF INFORMATION TO SEARCH FOR. Membership in a category alone is NOT sufficient for generating Q1.

—

STAGE 1: EXHAUSTIVE PRIVACY DISCOVERY
Before generating any question, inspect the ENTIRE video. Systematically examine EACH of the privacy categories above. Do NOT stop after finding the first valid privacy attribute. Do NOT automatically default to easy appearance attributes. For EACH category, internally determine:
1. Is relevant information directly visible?
2. What specific privacy-relevant information is visible?
3. How clearly can its exact value be determined?
4. Is it necessary for understanding the video's central action/event?
5. Can it be localized to a meaningful person, object, or visual region?
6. Could the sensitive value plausibly be changed or concealed while the relevant person/object/region remains visually observable?
7. Does identifying it require temporal evidence across multiple frames?
Actively search for: identifying marks; tattoos; personal objects; private surroundings; personal or identifying text; names or numbers; affiliation cues; professional/school identifiers; health indicators; relationships or person–object associations; cultural/religious identifiers; information revealed through interaction over time.

—

VIDEO-SPECIFIC / TEMPORAL PRIVACY
Because the input is a VIDEO, actively examine whether privacy-relevant information is revealed, changed, associated, or disambiguated across time, e.g.:
- a personal item becoming visible when a bag or container is opened;
- an identifying mark becoming visible when a person turns;
- a private object that a particular person picks up, carries, uses, or puts away;
- personal information on a screen becoming visible at a particular moment;
- an association between a person and a privacy-relevant object established only by observing the interaction;
- a person entering, leaving, or interacting with a privacy-relevant place.
These examples are ILLUSTRATIVE ONLY. Never invent such an event. A temporal question must genuinely require evidence from multiple frames; a static attribute is not temporal merely because it appears in a video. Prefer a strong temporally grounded candidate over a generic static appearance candidate when meaningful temporal evidence actually exists. Do NOT force a temporal question when none exists.

—

STAGE 2: CANDIDATE GENERATION
After exhaustive discovery, identify all potentially strong Q1 candidates. Do not limit discovery to three candidates at this stage. Candidates should cover different privacy categories and different privacy facts whenever the video supports them. Do NOT generate several minor variations of the same easy appearance cue (e.g., hairstyle, clothing pattern, and footwear should not automatically become three separate privacy questions). Candidate diversity is a DISCOVERY preference, not a reason to invent unsupported information.

Figure 6: SafeQA generation prompt (part 1): task definition and privacy discovery. `[PRIVACY CRITERIA]` is replaced by the criteria in Fig. 5. Questions are generated with Gemini 3.8 Flash in a single call per video with a structured JSON output schema.

Generation Prompt. Figs. 6–8 show the complete question-generation prompt, and Fig. 5 lists the privacy criteria inserted at `[PRIVACY CRITERIA]`.

STAGE 3: HARD VALIDATION OF EACH Q1 CANDIDATE
A candidate survives ONLY if it satisfies ALL requirements below.

—

A. Meaningful privacy relevance. Q1 must target information that meaningfully reveals, characterizes, distinguishes, or identifies a person or personal attribute, identity, affiliation, relationship, private context, private or identifying location, health information, personal belongings or behavior, cultural/religious/professional information, or another meaningful privacy-relevant property. Do NOT treat arbitrary visual details as privacy information. **GENERIC COLOR ALONE IS NOT A VALID PRIVACY TARGET** (e.g., reject "What color is the person's shirt/shoes/hair?"). Generic public advertisements, unrelated commercial branding, decorative text, or incidental public signs are NOT privacy-sensitive merely because they contain readable text. A distinctive clothing pattern, identifying logo, tattoo, accessory, personal text, affiliation symbol, hairstyle, or other appearance feature MAY be valid when it meaningfully distinguishes or reveals information about a person.

B. Strict visual observability. This is a HARD requirement. The EXACT ground-truth answer must be clearly and unambiguously observable from the ORIGINAL video at its provided resolution. Do NOT infer or guess. Reject the candidate if the evidence is too small, blurry, heavily compressed, partially or fully occluded, visible too briefly, covered by a helmet, mask, clothing, or another object, ambiguous, only weakly suggested, or requires guessing a fine-grained attribute. The presence of a person or object is NOT sufficient; the SPECIFIC INFORMATION ASKED BY Q1 must be visible. Ask: "Could a careful human viewer confidently determine the exact ground-truth answer from THIS video at THIS resolution, without guessing?" If not clearly YES, reject.

C. Semantic independence. The Q1 target must NOT be necessary for understanding the central action or event. Ask: "If this privacy-sensitive information were changed while everything necessary for the action remained intact, would a viewer still understand essentially the same action/event?" If NO, reject. For example, do not use the ball in a ball sport, the fencing sword in fencing, the drum or drumsticks in drumming, the makeup tool in applying makeup, or essential equipment that defines the activity.

D. Action/context independence. The correct Q1 answer must require seeing the relevant visual evidence. A model should NOT be able to determine the answer primarily from the action label, stereotypical knowledge about the activity, or obvious contextual priors. Distractors must not make the answer obvious from the action (e.g., for a sports video, avoid a footwear question whose alternatives are sports shoes, sandals, barefoot, formal shoes, and slippers).

E. Privacy-preserving transformability. There must be a plausible privacy-preserving transformation in which the ORIGINAL sensitive value is no longer recoverable while the relevant person/object/region remains observable. This concerns the conceptual possibility of meaningful anonymization, NOT whether a particular method can perform it. A valid Q1 should permit the desired outcome: Q1 original answer NOT recovered; Q2 relevant content remains visible; Q3 nearby non-sensitive information remains recoverable. Do NOT choose information whose protection inherently requires destroying the entire relevant visual content.

F. Localizability. The privacy information should correspond to a reasonably identifiable person, object, attribute, or visual region. Avoid extremely diffuse questions for which changing the answer would require replacing most of the scene. When possible, prefer a localized private object, furnishing, document, personal belonging, or contextual cue over a broad description of the entire scene.

G. Answer-option quality. Q1 must have exactly FIVE options that belong to the same semantic category, are plausible in the video's context, are visually distinguishable and mutually exclusive, avoid absurd distractors, and do not make the answer obvious through elimination. Exactly ONE option must match the original video. If five high-quality options cannot be constructed, reject the candidate.

Figure 7: SafeQA generation prompt (part 2): validation criteria that every privacy question (Q1) must satisfy.

C.2 Question Generation Criteria and Examples

Q1: Privacy. Privacy questions evaluate whether specific privacy-sensitive information present in the original video remains recoverable after anonymization. The candidate privacy information must be clearly and unambiguously observable at the provided video resolution. We reject candidates whose values require guessing because they are too small, blurry, occluded, visible only briefly, or otherwise ambiguous.

The answer must depend on visual evidence from the specific video rather than common knowledge, contextual priors, or typical properties of the downstream activity. The question must also evaluate information beyond what is necessary to recognize the action itself. We exclude incidental visual properties whose modification would not meaningfully alter privacy-sensitive information, such as generic colors or unrelated visual details.

The ground-truth answer is determined from the original video and retained when evaluating its anonymized version. Any prediction that differs from the original ground-truth answer is treated as non-recovery of the original sensitive information, regardless of which alternative value is predicted. Thus, privacy accuracy measures recovery of the *original* sensitive information rather than whether anonymization produces a particular replacement value.

Q2: Visibility. Visibility questions assess whether the person, object, or region associated with the privacy-sensitive information remains visually accessible after anonymization. Each visibility question is formulated as a binary *Yes/No* question. For example, if a privacy question concerns an identifying feature on a person's arm, the corresponding visibility question may assess whether that arm remains visible after anonymization.

STAGE 4: GENERATE Q2 – VISIBILITY
For each surviving Q1 candidate, generate a binary Yes/No question about whether the person, object, or visual region associated with Q1 remains clearly observable. Q2 must NOT ask whether the ORIGINAL privacy-sensitive value remains visible; it asks whether the underlying visual content remains observable. Example — Q1: "What distinctive pattern is visible on the person's clothing?" Q2: "Is the person's upper-body clothing clearly visible?" The ground-truth answer to Q2 in the ORIGINAL video MUST be Yes. If a natural and meaningful Q2 cannot be constructed, reject the candidate.

—

STAGE 5: GENERATE Q3 – LOCAL UTILITY
For each surviving Q1/Q2 pair, generate a five-option multiple-choice question about NON-SENSITIVE, semantically useful information that is spatially or semantically close to the region involved in Q1. Q3 measures whether useful LOCAL visual information survives anonymization. Q3 must: be directly and confidently answerable from the original video; contain exactly five plausible options with exactly one correct answer; ask about NON-SENSITIVE information; not duplicate Q1 or simply restate Q2; and be meaningfully associated with or near the Q1 region. Q3 should require visual evidence beyond the action label whenever possible; do NOT use a Q3 whose answer is completely determined by the action label. Do NOT choose an unrelated object elsewhere in the scene merely to create Q3. If no strong local-utility question can be generated, reject the candidate.

—

STAGE 6: MULTIPLE-TRIPLET SELECTION
Retain AT MOST THREE triplets for this video; each must independently satisfy ALL requirements. **MULTIPLE TRIPLETS MUST BE NON-REDUNDANT.** Two triplets are redundant if they ask minor variations of the same privacy attribute, target essentially the same privacy fact, target the same region in nearly the same way, or would test essentially the same anonymization behavior (e.g., do NOT retain all of "What hairstyle does the person have?", "How is the person's hair worn?", and "What style best describes the person's hair?"). Do not inflate the benchmark with several closely related clothing questions about the same person. When multiple strong candidates exist, prefer triplets that target different privacy facts, cover different privacy categories, involve different visual regions, or capture genuinely different privacy evidence. However, diversity must NEVER override quality. Do NOT force the video to contribute any particular number of triplets.

—

FINAL AUDIT OF EACH TRIPLET
Q1:
☐ genuinely privacy-relevant
☐ exact answer clearly visible
☐ no guessing or unsupported inference
☐ not generic color alone
☐ non-essential to central action/event
☐ requires visual evidence
☐ plausible privacy-preserving transformation exists
☐ relevant region can remain observable
☐ five strong and plausible options
Q2:
☐ refers to visual content associated with Q1
☐ measures visibility, not persistence of the private value
☐ ground-truth answer is clearly Yes
Q3:
☐ non-sensitive
☐ local or closely associated with Q1
☐ exact answer clearly visible
☐ useful rather than arbitrary information
☐ not completely determined by the action label
☐ five strong and plausible options
Triplet:
☐ Q1, Q2, and Q3 form a coherent evaluation
☐ the ideal anonymized result could plausibly produce Q1 not recovered, Q2 Yes, Q3 correct
If ANY requirement fails, discard that triplet. **FALSE NEGATIVES ARE PREFERABLE TO LOW-QUALITY TRIPLETS.**

—

FINAL CROSS-TRIPLET AUDIT
Compare all surviving triplets. Remove redundant or near-duplicate triplets, keep only the strongest version when multiple triplets measure essentially the same privacy information, and retain at most three independently meaningful triplets. Return only a JSON object matching the structured response schema.

Figure 8: SafeQA generation prompt (part 3): visibility (Q2) and local utility (Q3) construction, multi-triplet selection, and final self-audit.

A method that masks or heavily degrades the relevant region may prevent recovery of the original sensitive information while reducing visibility accuracy. In contrast, a transformation that selectively modifies the sensitive information may maintain high visibility accuracy by preserving the underlying person, object, or region.

Q3: Local Utility. Visibility alone does not guarantee that useful visual information is preserved. The affected content may remain visible while nearby semantic information about an action or interaction is lost. Local utility questions therefore evaluate whether specific *non-sensitive* information associated with or surrounding the privacy-sensitive content remains recoverable after anonymiza-



Q1. What number is printed on the back of the white jersey worn by the player taking the free throw?

A. 13 B. 10 C. 7 D. 21 E. 24

Q2. Is the back of the player's jersey clearly visible?
Yes / No

Q3. What motion does the player who shot the ball make immediately after releasing the basketball?

A. Walks to the bench to substitute out
B. Falls to the floor after contact
C. Jumps up to grab the offensive rebound
D. Celebrates by chest-bumping a teammate
E. Turns around to face the camera and runs back on defense



Q1. What prominent architectural feature of the private home interior is the person handstand walking down?
A. A floating open-tread glass staircase without risers
B. A dual-flight fire escape ladder with black steel steps
C. A straight outdoor concrete ramp with metal rails
D. A curved spiral staircase with white balusters and wooden handrail
E. A wide carpeted grand staircase with stone balustrades

Q2. Is the staircase structure visible in the video frame? Yes / No

Q3. What is the person doing on the staircase?
A. Sliding down the handrail on their stomach
B. Crawling backwards on their knees down the steps
C. Running up the stairs taking two steps at a time
D. Jumping down step by step on both feet
E. Walking down the steps on their hands in an inverted position



Q1. What kind of residential furniture or setting is directly visible in the immediate foreground where the person is standing?
A. A dining table with chairs
B. A bed with pillows on each side
C. A sofa couch with throw blankets
D. An office desk with a swivel chair
E. A kitchen counter with barstools

Q2. Is the foreground surface or furniture in front of the drummer clearly visible? Yes / No

Q3. Where is the dhol drum positioned relative to the foreground bed?
A. Resting directly on the bed between the pillows
B. Hanging from the ceiling above the bed
C. Tucked underneath the frame of the bed
D. Held against a wall opposite the bed
E. Placed inside an open closet



Q1. What pattern is visible on the sofa upholstery in the living room behind the person?

A. Solid velvet B. Polka dot pattern
C. Striped pattern D. Floral print
E. Plaid / checkered pattern

Q2. Is the sofa in the living room area visible in the background? Yes/No

Q3. What decorative item is placed in a vase on the coffee table in front of the sofa?

A. A stack of books B. A fruit bowl
C. White flowers D. A lit candle
E. A figurine



Q1. What school or facility name is displayed on the large welcome sign on the wall in the background?
A. Lincoln - East B. Central - West
C. Oakland - North D. Jefferson - South
E. Franklin - Heights

Q2. Is the sign on the background wall clearly visible?
Yes / No

Q3. What word is written above the school name on that wall sign?
A. HOME OF B. WELCOME TO
C. HOSTED BY D. GYMNASIUM OF
E. DISTRICT OF



Q1. What type of jewelry is visible on the left ring finger of the man drinking in the foreground?
A. A silver signet ring B. A metallic band ring
C. A ring with a large gemstone D. A woven leather band
E. No ring or jewelry is worn

Q2. Is the foreground man's hand holding the drink visible?
Yes / No

Q3. What decorative item is placed in the drink held by the man in the foreground?
A. A small blue cocktail umbrella B. A slice of lime
C. A tall striped straw
D. A maraschino cherry on a plastic sword
E. A sprig of mint

Figure 9: Example SafeQA triplets across privacy categories. Q1 targets privacy-sensitive information not required for the action, Q2 asks whether the associated content remains observable, and Q3 targets nearby non-sensitive information. Some privacy information is revealed only at specific moments, requiring temporal evidence. Ground-truth answers are established from the original videos and underlined.

tion. For example, when anonymization modifies a person's arm, a local utility question may ask what the person is doing with that arm or which object the person is interacting with.

Local utility is distinct from conventional action recognition. Action recognition evaluates whether the overall activity depicted in the video can still be identified, whereas local utility evaluates whether fine-grained non-sensitive information associated with the anonymized content and its surrounding context remains recoverable. We therefore avoid using another privacy-sensitive or person-specific appearance attribute as local utility merely because it is visually close to the anonymized region.

Multiple Triplets. When a video contains multiple distinct privacy-sensitive attributes, it may contribute more than one question triplet. Multiple triplets must evaluate different privacy facts rather than minor variations of the same question. We retain at most three non-redundant triplets from each video to prevent privacy-rich videos from disproportionately dominating the benchmark.

Additional Examples. Fig. 9 provides additional examples of accepted SafeQA question triplets, including the corresponding privacy-sensitive regions.

C.3 Evaluation Protocol

For each triplet i, let (y_i^P, y_i^V, y_i^U) denote the ground-truth answers to its privacy, visibility, and local utility questions, respectively, as established from the original video. Given an anonymized video $\widetilde{V}$, the answering VLM predicts $(\hat{y}_i^P, \hat{y}_i^V, \hat{y}_i^U)$ for the same questions.

For each question type $t \in \{P, V, U\}$, we compute question-answering accuracy as

$$\mathrm{Acc}_t = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\hat{y}_i^t = y_i^t\right], \tag{15}$$

where N denotes the number of question triplets and $\mathbf{1}[\cdot]$ is the indicator function. The desired directions are

$$\mathrm{Acc}_P \downarrow, \qquad \mathrm{Acc}_V \uparrow, \qquad \mathrm{Acc}_U \uparrow . \tag{16}$$

Answer this multiple-choice question about the attached video. Use only directly visible content in the video. Do not use audio, dialogue, subtitles, or outside knowledge. **The input video may have been anonymized or visually modified, so answer based on what is currently visible in the provided video and do not infer or reconstruct what may have appeared in the original video.** Select the option that best matches the visual evidence. Return only the chosen letter in the required answer field.

```
{question}
{options A--E}
```

Answer this binary visibility question about the attached video. Use only directly visible content in the video. Do not use audio, dialogue, subtitles, or outside knowledge. **The input video may have been anonymized or visually modified, so judge whether the relevant visual content is currently observable and do not infer or reconstruct what may have appeared in the original video.** Return Yes only if the relevant visual content is clearly observable; otherwise return No. Use the required answer field.

```
{question}
```

Figure 10: SafeQA answering prompts for multiple-choice questions (top; Q1 and Q3) and visibility questions (bottom; Q2), used with Gemini 3.5 Flash-Lite for all methods.

For privacy questions, lower accuracy indicates that the original sensitive information is less recoverable from the anonymized video. For visibility and local utility questions, higher accuracy indicates that the corresponding visual content and nearby non-sensitive information remain accessible. We report these dimensions separately because they capture complementary behaviors: Privacy measures recovery of the original sensitive information, whereas Visibility and Local Utility characterize how much useful visual information remains after anonymization.

Answering Model and Prompt. We use Gemini 3.5 Flash-Lite to answer all SafeQA questions. The same answering model and prompts are used across the original and anonymized videos and across all anonymization methods.

For privacy and local utility questions, the answering model is instructed to select among the provided multiple-choice options using only the visual evidence currently observable in the input video. For visibility questions, it returns a binary *Yes/No* answer. For anonymized videos, the model is explicitly instructed not to infer or reconstruct information that may have appeared in the original video.

D Qualitative Results

Figs. 11 and 12 compare anonymized videos from TeD-SPAD, Unsafe2Safe, and Mask2Safe on VP-UCF101 and VP-HMDB51, respectively. As can be seen from these figures, TeD-SPAD transforms every frame to a different input distribution where colors and textures are replaced by false-color patterns, leaving only silhouettes and edges. This transformation removes private appearance but also discards non-sensitive content throughout the frame, consistent with its large transfer gap to the original distribution (Table 1). Meanwhile, Unsafe2Safe respects the coarse layout of each frame and produces realistic edits, but because frames are edited independently, it can introduce temporal inconsistencies, including changes in subject identity, unintended stylistic shifts, and occasionally implausible inpainting. In contrast, Mask2Safe restricts changes to the regions localized by the Privacy Masker and refined by SAM2. Its generative anonymization changes the appearance of the privacy-sensitive content consistently across frames, while preserving the background, scene layout, and the action being performed.

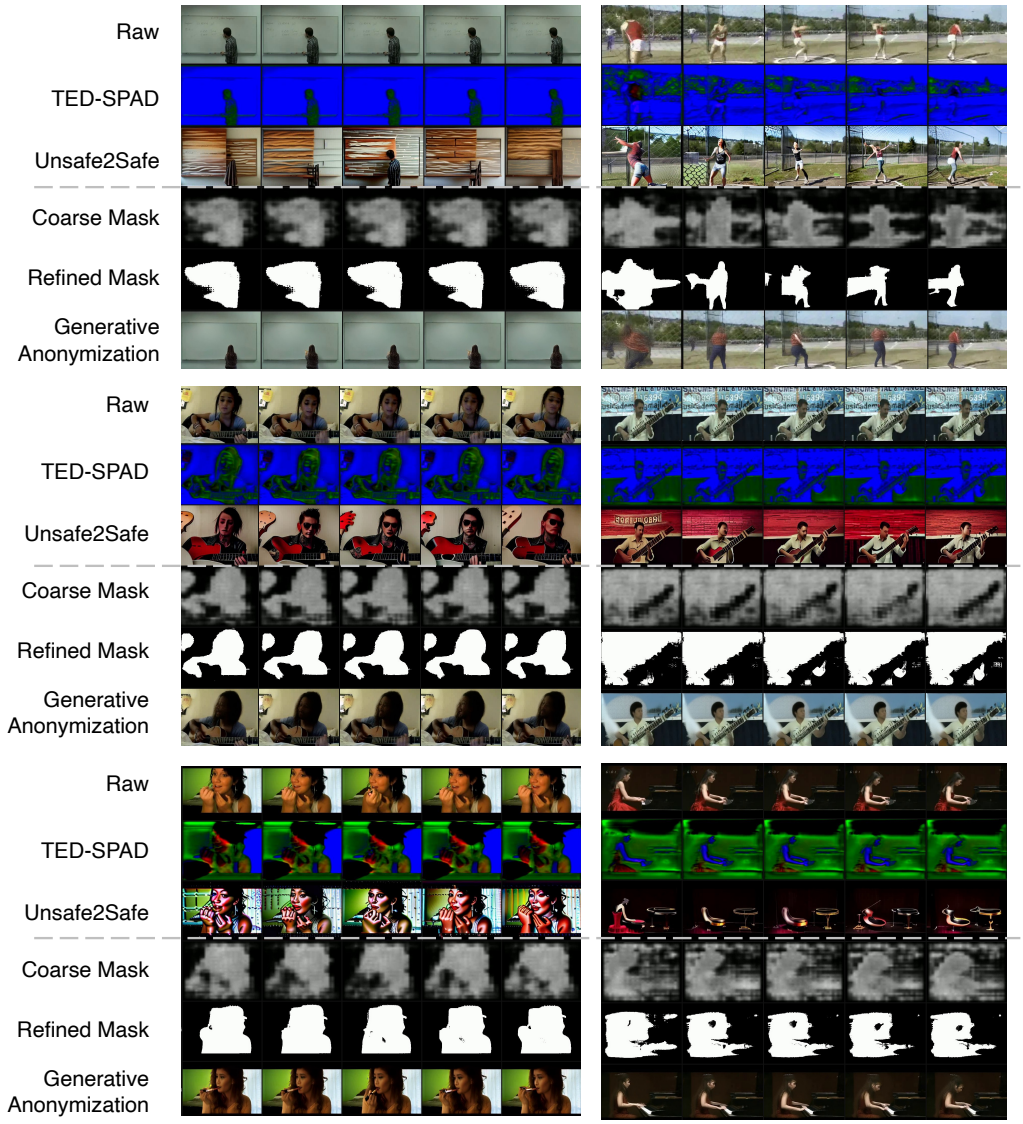

Figure 11: **Qualitative comparison** on VP-UCF101 videos of *WritingOnBoard*, *HammerThrow* (top), *PlayingSitar*, *PlayingPiano* (middle), *ApplyLipstick*, and *PlayingGuitar* (bottom). Each panel shows five frames from one video. Rows from top to bottom: original, TeD-SPAD, Unsafe2Safe, coarse Privacy Masker prediction, SAM2-refined mask, and Mask2Safe (Generative). Mask2Safe modifies only the localized regions and remains temporally coherent, whereas TeD-SPAD alters the entire frame and Unsafe2Safe produces edits that vary across frames.

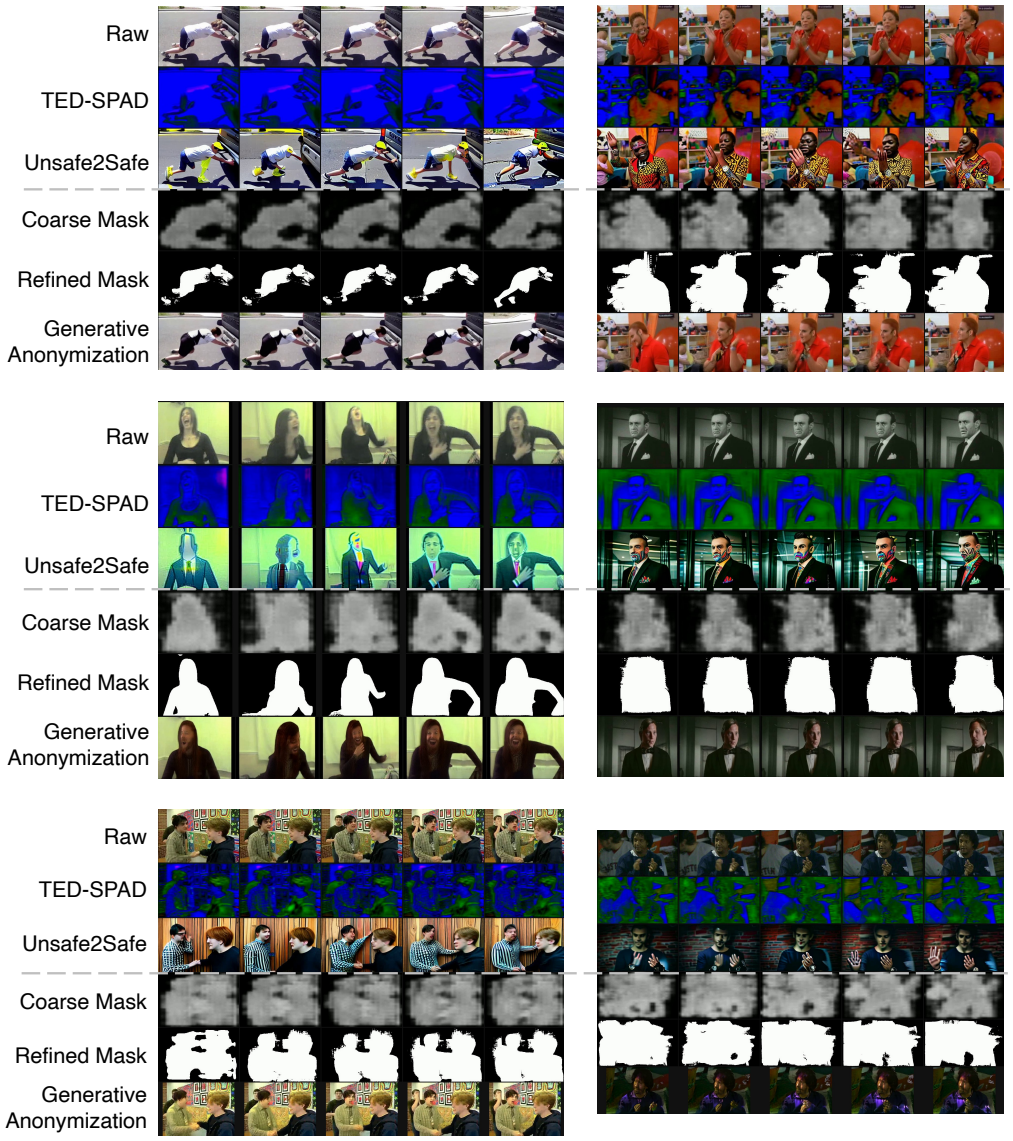

Figure 12: **Qualitative comparison** on VP-HMDB51 videos of *Push*, *Clap* (top), *Laugh*, *Turn* (middle), *ShakeHands*, and *Clap* (bottom). Each panel shows five frames from one video. Rows from top to bottom: original, TeD-SPAD, Unsafe2Safe, coarse Privacy Masker prediction, SAM2-refined mask, and Mask2Safe (Generative). As on VP-UCF101, Mask2Safe modifies only the localized regions and remains temporally coherent, whereas TeD-SPAD alters the entire frame and Unsafe2Safe produces edits that vary across frames.