Less Static, More Private: Towards Transferable Privacy-Preserving Action Recognition by Generative Decoupled Learning

Supplementary Material

The supplementary material is organized as follows:

- Sec. A: Dataset Details, including dataset statistics of our benchmarks and the privacy attributes definitions.
- Sec. B: Implementation Details, including hyperparameter settings, model architectures, and training procedures.
- Sec. C: Additional Experimental Results, including comprehensive ablation studies, same-domain evaluations, and a model complexity analysis.
- Sec. D: Qualitative Results, including feature distribution visualizations and anonymized frame examples.
- Sec. E: Visual Aid of Training and Evaluation, including diagrams illustrating the core stages of our framework.

A. Datasets

UCF ↔ HMDB. The *UCF↔HMDB* benchmark was released by Chen *et al*. [1] for studying video domain adaptation in action recognition. This transfer benchmark comprises 3,209 action videos spanning 12 action classes. The 12 overlapping classes of these two datasets are shown in Table S1. We use the official splits provided by Chen *et al*. [1]. The five human privacy attributes, *i.e*., *face, skin color, gender, nudity, and familiar relationship,* provided by Li *et al*. [10], are used for training the privacy attributes predictor during adversarial learning. This transfer benchmark, combining privacy attributes and overlapping action labels, is named the *TP-UCF↔TP-HMDB* benchmark.

UCF	HMDB
RockClimbingIndoor, RopeClimbing	climb
Fencing	fencing
GolfSwing	golf
SoccerPenalty	kick_ball
PullUps	pullup
Punch, Boxing(Punching/Speed)Bag	punch
PushUps	pushup
Biking	ride_bike
HorseRiding	ride_horse
Basketball	shoot_ball
Archery	shoot_bow
WalkingWithDog	walk

Table S1. The overlapping action classes shared by UCF and HMDB in the *UCF ↔ HMDB* benchmark.

NEC-Drone ↔ Kinetics. This transfer benchmark, composed of two datasets: *NEC-Drone* [2] and *Kinetics* [8], comprises videos of 7 overlapping classes. The overlapping label sets between *NEC-Drone* and *Kinetics* are listed in Table S2. Our splits are based on the official train/test splits provided by Choi *et al*. [2]. The Privacy-preserving action recognition experiments in our work are conducted in both *Kinetics→NEC-Drone* and *NEC-Drone→Kinetics* settings, which is more challenging than *UCF ↔ HMDB*, as the domain gap between source and target domains is more significant due to the viewpoint variations introduced by drone-captured footage.

NEC-Drone	Kinetics
walking	marching
running	jogging, running on treadmill
jumping	high jump, jumping into pool
drinking water from a bottle	drinking beer
throwing an object	throwing axe, throwing ball, throwing discus, shot put, javelin throw
shaking hands	shaking hands
hugging	hugging

Table S2. Correspondences between action classes of *NEC-Drone* and *Kinetics* for the same label set in source and target settings.

ARID ↔ HMDB. The *ARID ↔ HMDB* benchmark consists of videos spanning 11 overlapping classes, selected from *ARID*[17] and *HMDB*[9] datasets. These overlapping classes include the following actions, *i.e*., *Drink, Jump, Pick, Pour, Push, Run, Sit, Stand, Turn, Walk, and Wave.* The splits are based on the official train/test partition provided by Xu *et al*. [17]. In this benchmark, we conduct privacy-preserving action recognition experiments considering both *ARID→HMDB* and *HMDB→ARID* directions.

VISPR. We conduct cross-dataset training and evaluation using *VISPR*[12] dataset. In our cross-dataset training and evaluation setting, we utilize 7 privacy attributes commonly present in the action video datasets mentioned above, as selected and provided by Wu *et al*. [16]. These attributes include *semi-nudity, occupation, hobbies, sports, personal relationships, social relationships, and safety*. The multi-attribute predictor is trained on *VISPR* to support adversarial training and privacy-preserving evaluation.

Jester. We additionally conduct cross-dataset evaluation on *Jester*$\mathcal{S}$ ↔ *Jester*$\mathcal{T}$, a cross-domain and fine-grained hand gesture recognition benchmark. Constructed by Lin *et al*. [11] from the *Jester* dataset, this benchmark defines transfer tasks over seven gesture classes, with the source

and target domains containing 51,498 and 51,415 video clips, respectively. The primary challenge of this benchmark lies in leveraging temporal dynamics over solely spatial cues for accurate recognition. For the experiments in Sec C.5, we used a 10% subset of the official training and testing data in a class-balanced manner to validate our method's transfer performance on temporal-dominant data.

B. Implementation Details

B.1. Architectural details of our anonymizer.

We implement our anonymizer f_A adopting a *VQGAN* video generation framework inspired by Ge *et al*. [7]. In the pretraining process of f_A, we follow [5] to stabilize the training of the quantizer and its associated codebook, ensuring that the updated embeddings in the codebook are close to the selected codebook while preventing the encoder output from deviating from the codebook. To improve the reconstruction quality, our reconstruction loss also incorporates perceptual and discrimination terms, a standard practice in generative models [5, 7]. We employ a quantized encoder as it significantly improves domain transfer performance. The results in Table S6 suggest that our vector quantization constraint aids in learning robust action semantics, thereby improving the cross-domain transfer performance of the privacy-preserving action recognition model.

B.2. Implementation details for loss functions.

When calculating the *Spatial Consistency Loss*, $\mathcal{L}_{\text{s-cons}}$, a projector is used to map the features of the anchor, positive, and negative samples to the same dimensional space. We employ a domain projector with a Gradient Reversal Layer (*GRL*), a technique well-established by DANN [6], and a domain classifier for dynamic feature extraction and calculating the *Temporal Alignment Loss*, $\mathcal{L}_{\text{t-align}}$. The *GRL* reverses the gradient's direction during backpropagation. This trains the encoder to generate dynamic features that can confuse the domain classifier, thus making the extracted dynamic features domain-invariant. This adversarial process, optimized via our *Temporal Alignment Loss*, $\mathcal{L}_{\text{t-align}}$, results in robust dynamic representations that are aligned across the source and target domains.

B.3. Training algorithm of our framework.

Let's consider the models f_A, f_T, f_B, parameterized by θ_A, θ_T, and θ_B, respectively. $\mathbb{D}_{\mathcal{S}}$ is the source domain dataset and $\mathbb{D}_{\mathcal{T}}$ is the target domain dataset. We summarize the entire training process of our framework in algorithm 29.

B.4. Hyperparameter selection details.

To balance the impact of the action recognition loss, ($\mathcal{L}_{\text{act}}$) and privacy prediction loss, ($\mathcal{L}_{\text{pri}}$) when updating the anonymizer during adversarial training, we set the weights

Algorithm 1: GenPriv Framework

Inputs:
  Source Dataset: $\mathbb{D}_{\mathcal{S}} = \{(\mathbf{V}_i^{\mathcal{S}}, Y_i^{\mathcal{S}}, P_i^{\mathcal{S}})\}_{i=1}^{N_{\mathcal{S}}}$;
  // Labeled Action and Privacy
  Target Dataset: $\mathbb{D}_{\mathcal{T}} = \{\mathbf{V}_i^{\mathcal{T}}\}_{i=1}^{N_{\mathcal{T}}}$;
  // Unlabeled
  Epochs: max_init_epoch, max_adv_epoch
  Learning Rates: α_A, α_T, α_B
  Hyperparameters: λ_{st}, λ_a, λ_p, ω
Output:
  Anonymizer f_A^*, Action Recognizer f_T^*

Initialization:
  Initialize θ_T with Kinetics400 weights.
  Initialize θ_B with ImageNet weights.
  Initialize θ_A as an identity function ;
  // Encoder $f_{\mathcal{E}}$, Decoder $f_{\mathcal{G}}$, Codebooks $\mathcal{C}_F, \mathcal{C}_T$
for $e_{init} \leftarrow 1$ ***to*** *max_init_epoch* **do**
  Sample batch from $\mathbb{D}_{\mathcal{S}}$
  $\theta_A \leftarrow \theta_A - \alpha_A \nabla_{\theta_A}(\mathcal{L}_{rec}(\theta_A) + \mathcal{L}_{act}(\theta_A, \theta_T))$
end

Adversarial Learning:
for $e_{anon} \leftarrow 1$ ***to*** *max_adv_epoch* **do**
  Sample batch $b_{\mathcal{S}}$ from $\mathbb{D}_{\mathcal{S}}$ and $b_{\mathcal{T}}$ from $\mathbb{D}_{\mathcal{T}}$
  Step 1 Freeze θ_T, Freeze θ_B
  $\theta_A \leftarrow \theta_A - \alpha_A \nabla_{\theta_A} \Big[\lambda_{st}(\mathcal{L}_{mi} + \mathcal{L}_{s\text{-}cons} + \mathcal{L}_{t\text{-}align})$
  $+ \lambda_a \mathcal{L}_{act}(\theta_A, \theta_T) - \lambda_p \mathcal{L}_{pri}(\theta_A, \theta_B)\Big]$
  Step 2 Freeze θ_A
  $\theta_T \leftarrow \theta_T - \alpha_T \nabla_{\theta_T} \mathcal{L}_{act}(\theta_T, \theta_A)$
  $\theta_B \leftarrow \theta_B - \alpha_B \nabla_{\theta_B} \mathcal{L}_{pri}(\theta_B, \theta_A)$
end
return f_A^*, f_T^*

for the $\mathcal{L}_{\text{act}}$ and $\mathcal{L}_{\text{pri}}$ losses λ_{a} and λ_{t} to 1.0 and 0.5, respectively. The weight for spatial consistency and temporal alignment losses, λ_{st}, is set to 1.0. To ensure the realism of the generated videos during anonymizer initialization training, we initialize the *GAN* discriminator from the 50th epoch. To avoid overly aggressive transformations on the videos, we apply the reconstruction loss every 5 epochs. The Mutual Information, Static Consistency, and Temporal Alignment losses are introduced from the 10th, 15th, and 25th epochs, respectively. We set both the dynamic and static embedding dimensions to 256, with the static codebook containing 2,048 entries and the dynamic codebook comprising 16,384 embeddings. The initialization process is trained for 150 epochs, followed by 200 epochs of adversarial training, with the privacy attributes predictor retrained for evaluation for 50 epochs.

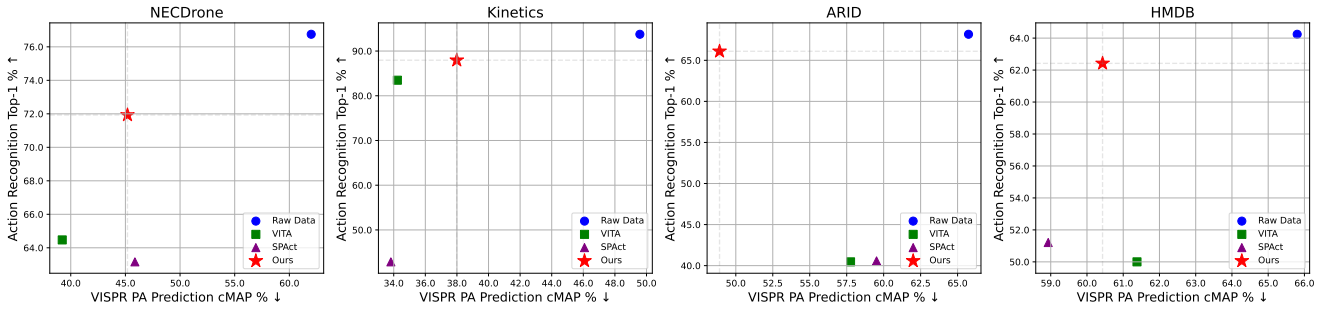

Figure S1. Same domain training and evaluation results on *V-Kinetics-NECDrone* and *V-ARID-HMDB* Cross-Dataset Benchmarks.

B.5. Domain adaptive action recognizer training.

For our action recognizer f_T, we use the R2plus1D-18 architecture [14], pre-trained on Kinetics400 [8]. f_T is initialized using standard domain adaptation methods [1, 3, 6, 13]. Source and target domain features are aligned by maximizing the domain classification loss, which aims to improve the action recognition model's transferability. High-confidence pseudo-labels are then used to augment the training data. Finally, an entropy loss is applied to the classifier outputs, encouraging a more peaked probability distribution and enhancing domain adaptation. Following the traditional unsupervised domain adaptation setting [1, 6], unlabeled target domain videos are randomly selected, with a batch size equal to that of labeled source domain data in each training iteration.

C. Additional Experimental Results

C.1. Same domain training and evaluation results.

To verify that our privacy-preserving module is not solely designed to achieve high action recognition performance in the transfer setting, we also need to ensure that it maintains strong action recognition performance in the same domain training and evaluation setting.

Results on TP-UCF↔TP-HMDB benchmark. Table S3 presents the Top-1 action recognition accuracy and privacy prediction results, including cMAP and F1 scores on TP-UCF↔TP-HMDB benchmark trained and evaluated on the source domain. The results indicate that our method effectively maintains the action-privacy trade-off in the transfer setting and the same domain training and evaluation setting.

Results on V-Kinetics-NECDrone and V-ARID-HMDB cross-dataset benchmarks. The plots in Figure S1 show the same domain training and evaluation results on V-Kinetics-NECDrone and V-ARID-HMDB Cross-Dataset Benchmarks, respectively, demonstrating that the performance of our model is comparable to existing PPAR methods under the same domain training and evaluation setting.

Method	TP-UCF			TP-HMDB		
	Top-1 (↑)	cMAP (↓)	F1 (↓)	Top-1 (↑)	cMAP (↓)	F1 (↓)
Raw data	98.24	72.51	59.19	94.16	68.47	55.23
VITA[16]	95.45	70.3	56.43	81.67	63.28	45.92
SPAct[4]	95.27	67.73	56.62	88.89	65.26	53.82
Ours	94.57	68.71	55.12	88.01	63.82	49.89

Table S3. **Same domain training and evaluation results on our TP-UCF↔TP-HMDB benchmark.** Comparison of PPAR models trained and evaluated on the same domain. Our method performs competitive action recognition while maintaining effective privacy preservation, comparable to existing approaches.

C.2. Model efficiency and complexity

We conducted a comparative analysis of model complexity to demonstrate the efficiency of our proposed framework and show that our performance gains are not solely due to model scale. We introduce a smaller variant, *GenPriv-small*, by reducing key parameters such as hidden feature dimensions. As shown in Table S4, this lightweight version still achieves a superior trade-off between task utility (Top1 accuracy) and computational cost (FLOPs) compared to prior privacy-preserving action recognition methods.

Method	Params.	FLOPs	$\text{Top1}_{H\to U}$(↑)	$\text{cMAP}_{H\to U}$(↓)	$\text{Top1}_{U\to H}$(↑)	$\text{cMAP}_{U\to H}$(↓)
VITA [38]	**1.3M**	166.0G	65.50%	70.60%	72.73%	64.25%
GenPriv-small	5.95M	**111.11G**	85.99%	68.36%	76.11%	65.37%
SPAct [9]	17.27M	160.32G	72.22%	66.47%	74.26%	64.89%
GenPriv	23.78M	406.71G	**87.91**%	67.42%	**80.55**%	64.84%

Table S4. **Model efficiency and performance comparison**. Our lightweight variant maintains a competitive utility-privacy balance with a notably lower computational cost (FLOPs).

C.3. Robustness of our anonymization function.

To further assess the robustness of our anonymization function, we evaluate the f_A^* on *TP-UCF↔TP-HMDB* using a *R3D* action classifier and a *R2plus1D* privacy attributes predictor, both distinct from those used during training. As shown in Table S5, *GenPriv* can still significantly reduce the cMAP to 67.22% on the *TP-HMDB→TP-UCF* setting, even with a novel video-based privacy attributes predictor.

This indicates that *GenPriv* effectively removes privacy-sensitive information from the entire video.

Method	TP-HMDB → TP-UCF			TP-UCF → TP-HMDB		
	Top-1 (↑)	cMAP (↓)	F1 (↓)	Top-1 (↑)	cMAP (↓)	F1 (↓)
Raw Video	86.16	72.31	58.70	80.83	67.41	54.92
VITA	83.32	71.88	55.97	68.06	66.03	52.43
SPAct	70.68	71.43	57.29	68.31	66.76	54.14
Ours	**84.94**	**67.22**	**52.79**	**69.72**	**65.77**	**49.78**

Table S5. Evaluate our f_A^* on unseen action and privacy models.

C.4. Analysis on the architecture designs.

To validate the design choice of our *ST-VAE* architecture, we conducted an ablation study comparing it against simpler alternatives. Our core innovation lies in the specific integration of ST-VAE within our generative decoupled learning framework. For a fair comparison, we adapted simpler *Autoencoder* (AE) and *Variational Autoencoder* (VAE) architectures to operate within our proposed framework. The results in Table S6 show that these alternatives struggle to achieve a good utility-privacy trade-off, while our *ST-VAE* based method achieves significantly superior performance.

Method	TP-HMDB → TP-UCF			TP-UCF → TP-HMDB		
	Top1↑	cMAP↓	F1↓	Top1↑	cMAP↓	F1↓
Source Only	85.81	72.51	0.592	78.69	68.47	0.552
AE+triplet	61.65	66.7	0.521	53.06	63.19	0.509
VAE+triplet	65.0	66.04	0.513	63.61	62.07	0.503
Ours(ST-VAE)	**87.91**	67.42	0.519	**80.55**	64.84	0.527

Table S6. **Ablation study on model architecture**. Compared to simpler alternatives like *AE* and *VAE*, our *ST-VAE* achieves a substantially better utility-privacy trade-off.

C.5. Experiments on other action video datasets.

To demonstrate the robustness and generalizability of our method on tasks that are more reliant on temporal dynamics, we further validated our approach on the Jester-DG/DA benchmark [13]. It is focused on fine-grained gesture recognition, requiring a strong understanding of temporal information. As shown in Table S7, our method demonstrates superior transfer performance, which is attributed to GenPriv's ability to preserve and align action-related dynamic features across domains.

Method	JesterS → JesterT			JesterT → JesterS		
	Top1↑	cMAP↓	F1↓	Top1↑	cMAP↓	F1↓
Source Only	58.10	65.72	0.571	62.41	65.14	0.573
VITA[38]	44.76	63.36	0.556	41.91	61.69	0.545
SPAct[9]	45.24	62.02	0.545	42.38	61.93	0.531
Ours	**49.87**	**60.87**	**0.441**	**53.81**	**57.17**	**0.472**

Table S7. Performance on the temporal-dominant Jester benchmark.

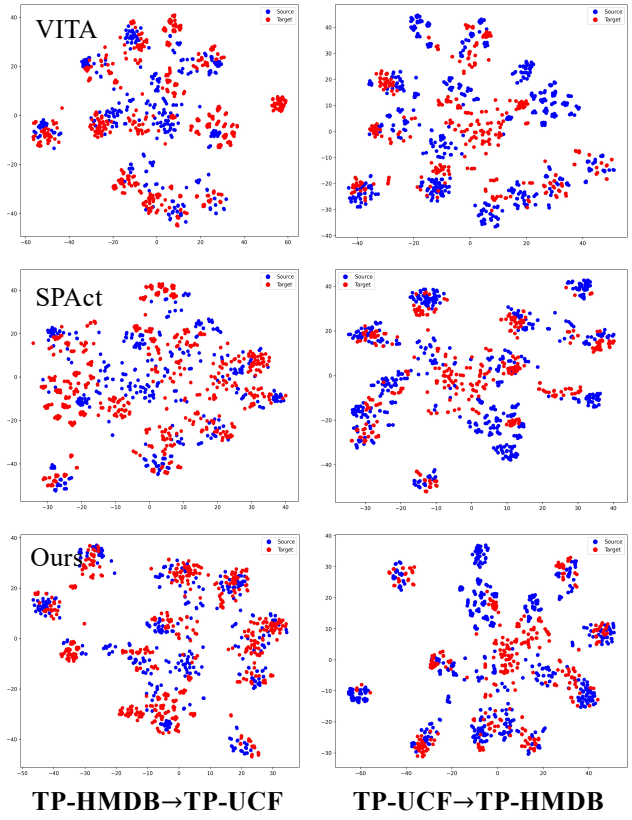

Figure S2. The comparison of *t-SNE* visualization with source (blue) and target (red) distributions on **TP-HMDB** ↔ **TP-UCF**.

C.6. Evaluation on diverse target domains

To further assess the generalization capabilities of our framework, we conducted an exploratory multi-target domain transfer experiment. This more challenging setting tests the model's ability to generalize from a single source domain to two different target domains simultaneously, providing a more rigorous evaluation of its robustness to diverse domain shifts. Table S8 shows that GenPriv exhibits superior generalization ability in action recognition across multiple targets while maintaining comparable privacy-preserving performance.

Method	$\mathcal{S}_{ucf} \rightarrow \mathcal{T}_{hmdb}$		$\mathcal{S}_{ucf} \rightarrow \mathcal{T}_{kinetics}$		$\mathcal{S}_{hmdb} \rightarrow \mathcal{T}_{ucf}$		$\mathcal{S}_{hmdb} \rightarrow \mathcal{T}_{kinetics}$	
	Top1↑	cMAP↓	Top1↑	cMAP↓	Top1↑	cMAP↓	Top1↑	cMAP↓
VITA[38]	71.81	63.32	76.60	55.62	67.37	70.65	75.94	57.76
SPAct[9]	65.25	64.44	70.09	63.72	63.39	67.87	73.21	61.75
[gray]0.95 **Ours**	**80.55**	64.84	**81.25**	60.71	**87.91**	67.42	**79.46**	58.56

Table S8. **Multi-target domain adaptation results**. Our GenPriv shows strong generalization from a single source ($\mathcal{S}$) domain to multiple, diverse target domains ($\mathcal{T}$).

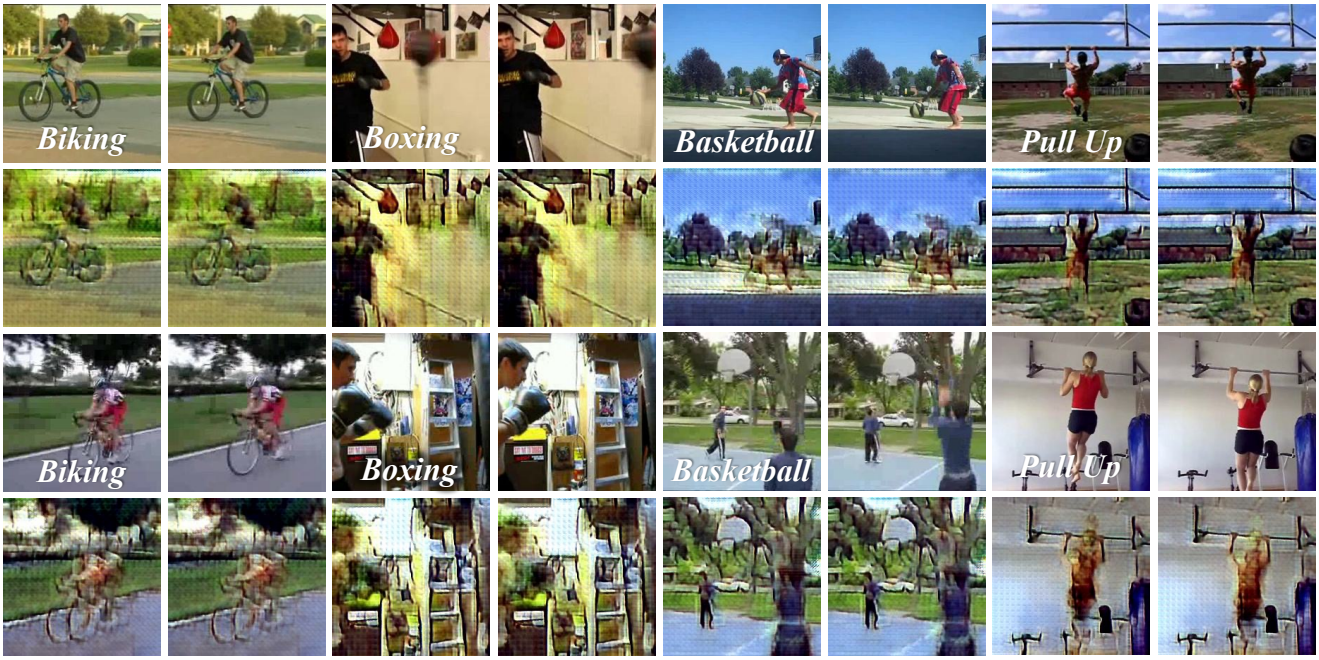

Figure S3. **Anonymized Frame Visualizations with GenPriv.** Examples of anonymized video frames are presented for *Biking, Basketball, Pull Up, and Boxing*, with source domain (top two rows) and target domain (bottom two rows). Every video shows 2 frames.

D. Additional qualitative results

Visualization of distribution. To verify that the action recognition model optimized during the adversarial process achieves stronger transferability on the target domain compared to other models, we visualize the distribution of both domains using t-SNE[15] to investigate how our approaches bridge the gap between the source and target domains. Specifically, we visualize the high-dimensional video features extracted before the last fully connected (FC) layer of the trained action recognition model. As shown in Figure S2, the features extracted by our method from the source and target domains achieve better alignment compared to the other two methods. This provides qualitative evidence of the transferability of our GenPriv framework.

Visualization of anonymized frames. Here, we present visualizations of video frames transformed by our GenPriv. As shown in Figure S3, our approach not only effectively protects privacy-sensitive attributes but also significantly preserves appearance and background information, which are crucial for accurate action recognition.

E. Visual Aid of Training and Evaluation

This section visually details the three core phases of our training and evaluation framework, as illustrated in Figures S4, S5, and S6.

- First, the adversarial training phase (Fig. S4): In this stage, the anonymization module is trained with both source and target domain videos. This stage uses source domain action and privacy labels to ensure action-relevant information is preserved while privacy-sensitive content is removed, ultimately yielding the learned anonymization function f_A^* and action recognition model f_T^*.
- Second, the privacy evaluation phase (Fig. S5): This phase consists of retraining a privacy attributes predictor, f_B^*, on the anonymized videos. To ensure a fair assessment, this retraining is conducted using the same privacy labels as the original training, allowing for a direct comparison of the predictor's performance before and after anonymization.
- Finally, the utility and budget evaluation phase (Fig. S6): This stage measures the final performance[cite: 951]. In this stage, the action utility and privacy budget are calculated on anonymized target domain evaluation videos using the trained action model f_T^* and the retrained privacy predictor f_B^*.

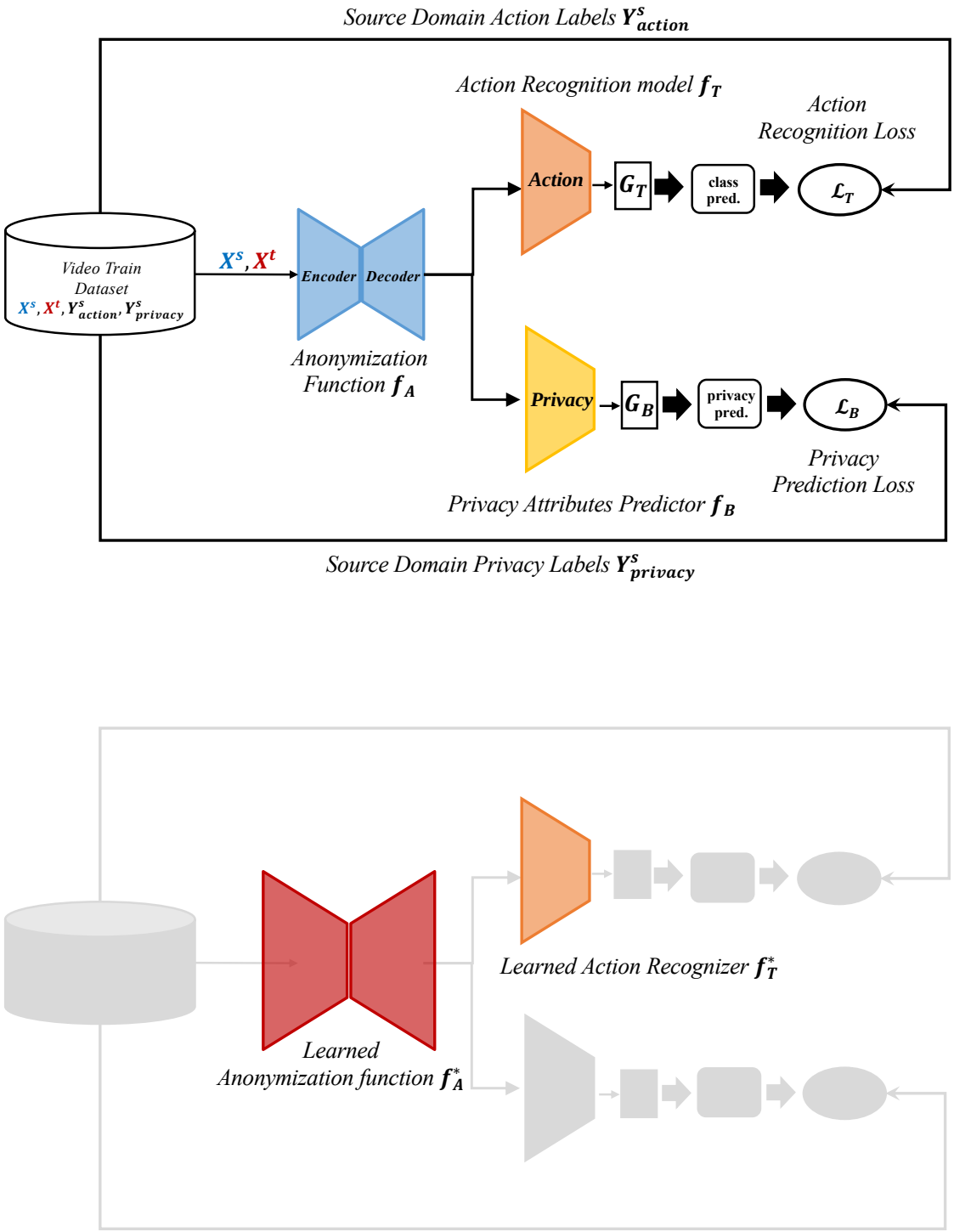

Figure S4. **(a) First phase: Adversarial Training.** In the adversarial training phase, the anonymization module is trained with source and target domain videos. Source domain action labels and privacy labels are used to ensure the reserving of action-relevant information while removing privacy-sensitive information, ultimately yielding the learned anonymization function f^*_A and action recognition model f^*_T.

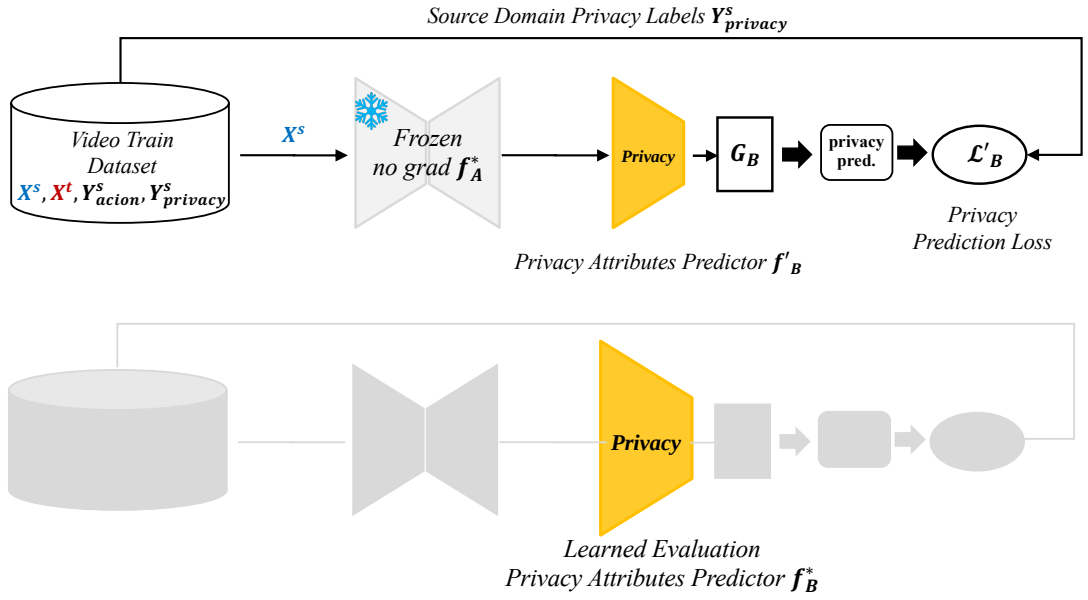

Figure S5. **(b) Second phase: Retrain for privacy evaluation.** Retrain a privacy attributes predictor f^*_B on the anonymized videos for privacy evaluation. The retraining is conducted using the same privacy labels as the original training, ensuring a fair comparison of the predictor's performance before and after anonymization.

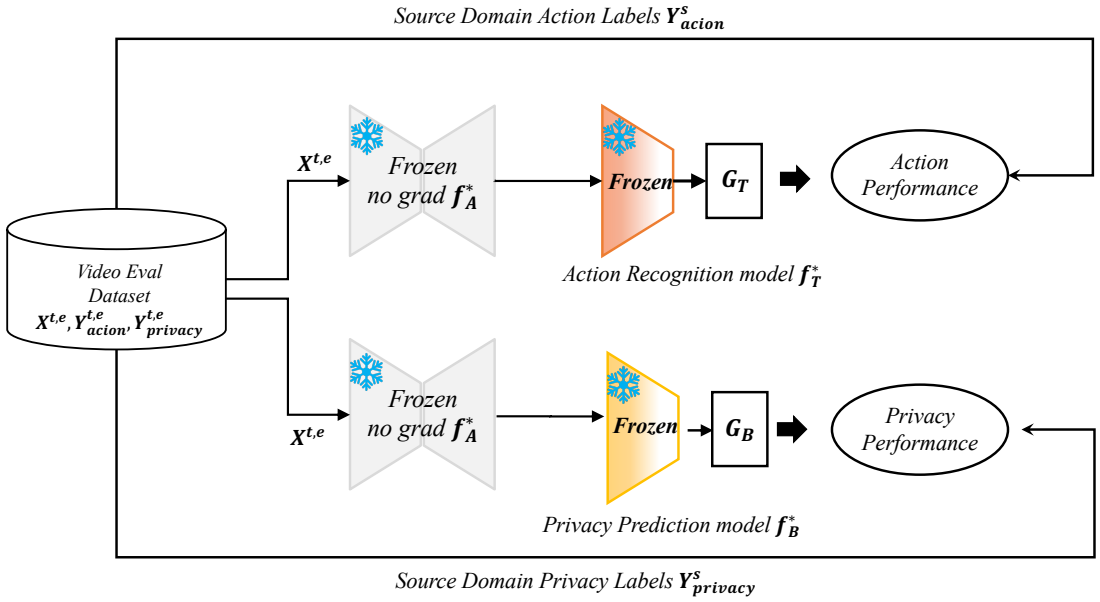

Figure S6. **(c) Third phase: Utility and Budget Evaluation.** In the evaluation stage, action utility and privacy budget are measured on anonymized target domain evaluation videos $X^{t,e}$ using the action recognition model f^*_T and the privacy attributes predictor f^*_B, supervised by the action labels $Y^{t,e}_{action}$ and privacy labels $Y^{t,e}_{privacy}$ (used only for evaluation).

References

[1] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *IEEE/CVF International Conference on Computer Vision*, pages 6321–6330, 2019. 1, 3

[2] Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1717–1726, 2020. 1

[3] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In *European Conference on Computer Vision*, pages 678–695. Springer, 2020. 3

[4] 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*, pages 20164–20173, 2022. 3

[5] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2

[6] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016. 2, 3

[7] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision*, pages 102–118. Springer, 2022. 2

[8] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The Kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 3

[9] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 1

[10] 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)*, pages 5106–5115, 2023. 1

[11] Kun-Yu Lin, Jia-Run Du, Yipeng Gao, Jiaming Zhou, and Wei-Shi Zheng. Diversifying spatial-temporal perception for video domain generalization. *Advances in Neural Information Processing Systems*, 36:56012–56026, 2023. 1

[12] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[13] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11815–11822, 2020. 3, 4

[14] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 3

[15] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 5

[16] 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 TPAMI*, 44(4):2126–2139, 2020. 1, 3

[17] Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiong Yin, and Simon See. ARID: A comprehensive study on recognizing actions in the dark and a new benchmark dataset. *CoRR*, abs/2006.03876, 2020. 1