When Visual Privacy Protection Meets Multimodal Large Language Models (Supplementary Material)

Xiaofei Hui[1], Qian Wu[2], Haoxuan Qu[1], Majid Mirmehdi[3], Hossein Rahmani[1], Jun Liu[1*]

[1*]School of Computing and Communications, Lancaster University, South Dr,, Lancaster, LA1 4WA, United Kingdom.
[2]Tongji University, Siping Road, Shanghai, 200092, China.
[3]Department of Computer Science, University of Bristol, Queens Road, Bristol, BS8 1QU, United Kingdom.

*Corresponding author(s). E-mail(s): j.liu81@lancaster.ac.uk;
Contributing authors: x.hui@lancaster.ac.uk; fivethousand@tongji.edu.cn;
h.qu5@lancaster.ac.uk; m.mirmehdi@bristol.ac.uk; h.rahmani@lancaster.ac.uk;

Appendix A More Ablation Studies

For a more comprehensive evaluation of our proposed framework, we conduct more ablation studies on UCF101-VISPR following [1, 2].

Impact of critical history gradients in the training process. As we consider the MLLM as a black-box function, in our framework, we estimate the gradient with zeroth-order optimization (SPSA algorithm). To facilitate the estimation, we propose to incorporate critical history gradients to tackle the randomness involved in the estimation. Here, to investigate the impact of the proposed critical history gradients in training, we record the variance of the estimated gradients along the optimization. We compare our method with the variant using only SPSA (i.e., without critical history gradient design) in Fig. A1. As can be observed, by applying the proposed critical history gradients, the variance of the estimated gradients tends to be much lower, implying that the randomness in the estimated gradients can be reduced.

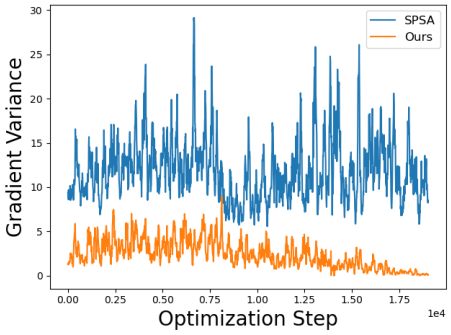

Fig. A1 Gradient variance during training process.

Impact of different weights λ_1 and λ_2. In Eq. 4 in the main paper, the weights λ_1 and λ_2 are applied to balance the two objectives. To evaluate our design under different weights, we conduct experiments with different values of λ_1 and λ_2. To further test the effectiveness of the designed overall objective, we also conduct experiments with the variant without this design, i.e., the variant that simply combines the objectives linearly using weighted sum objective (Eq. 2 in the main paper) under different weights. We compare the results

and show the trade-off plot in Fig. A2 and the numerical results in Tab. A1. As shown, all our results (for all varying weights) outperform the best results of the weighted sum objective in terms of both action recognition performance and privacy protection performance, demonstrating the effectiveness of our design.

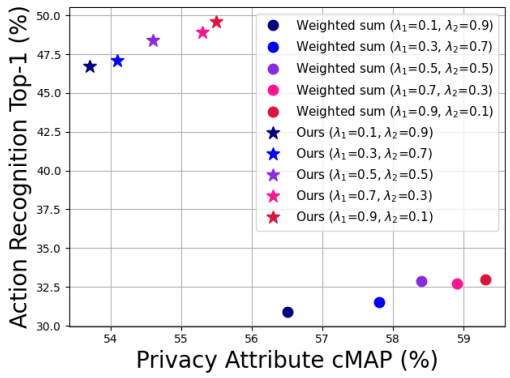

Fig. A2 Trade-off between action recognition and privacy preservation using different λ_1 and λ_2. **Note that higher value of action Top-1 and lower value of privacy cMAP indicate better performance, i.e., the performance tends to get better when the point in the figure gets closer to the *top left* corner. Best viewed in color.**

Impact of different values of γ. We adjust the direction of the estimated gradient in each step with history gradient information and control its impact with the factor γ (Eq. 7 in the main paper). Here, we conduct experiments with different values of γ ($\gamma = \{0, 0.3, 0.5, 0.7, 1\}$) to evaluate its impact. As shown in Tab. A2, when $\gamma = 0.5$, our method achieves the best performance. Thus, we set $\gamma = 0.5$.

Impact of different values of b. We adjust the step size in each optimization step and control its impact with the factor b in Eq. 8 in the main paper, where a larger value of b indicates a greater adjustment in step size. To evaluate the impact of different values of b, we conduct experiments with $b = \{0, 0.1, 0.3, 0.5, 1, 2\}$ and report the results in Tab. A3. Specifically, when $b = 0$, we do not adjust the step size. As shown, when $b = 0.3$, our method achieves the best performance. Thus, in our experiments, we set $b = 0.3$.

Impact of different values of m. We construct the critical gradient collection set with capacity m (i.e., containing m history gradients). Here, we conduct experiments with varying values of m to test its impact. We set $m = \{0, 5, 10, 20\}$ and report the results and the corresponding GPU memory consumption in Tab. A4. Specifically, when $m = 0$, no history gradient is collected. As can be seen, the performance increases when m increases. As the improvement of performance becomes trivial when m exceeds 10, out of consideration of efficiency, we set $m = 10$ in our experiments. Note that our method achieves significantly better performance compared with no history gradients ($m = 0$) with a small increase in GPU memory (around 1.1%).

Impact of different values of d. In the selection of critical history gradients, we set a decay factor d to reduce the impact of previous gradients over time (Eq. 6 in the main paper). Here, we evaluate the impact of different values of d and conduct experiments with $d = \{0.1, 0.5, 0.9, 1\}$. Specifically, when $d = 1$, the importance score s_{t_i} of the history gradient does not reduce over time, i.e., $s_{t_i} = ||\hat{g}_{t_i}||_2$. We show the results in Tab. A5. As shown, our method achieves the best performance at $d = 0.5$. Thus, we set $d = 0.5$ in our experiments.

Further analysis on the critical-history enhanced optimization. To better understand how our critical-history enhanced optimization works and why it achieves superior performance with fewer MLLM calls, we conduct analyses comparing the properties of selected and discarded gradients throughout the optimization process. Specifically, we compare the L2 norm distributions of selected versus discarded gradients across the entire optimization process. The results show that selected gradients have larger norms (mean $\pm$ std: 10.02 ± 8.61, median: 5.36) compared to discarded gradients (mean $\pm$ std: 6.24 ± 9.16, median: 2.04). As gradient norm is widely recognized as a key indicator of the gradient's contribution to optimization [3, 4], with larger norms indicating stronger and more informative descent directions, this shows that our selection mechanism effectively identifies gradients with stronger descent signals.

Table A1 Impact of different weights λ_1 and λ_2.

Method	UCF101-VISPR	
	Action (Top-1 ↑)	Privacy (cMAP ↓)
Weighted sum (λ_1=0.1, λ_2=0.9)	30.9	56.5
Weighted sum (λ_1=0.3, λ_2=0.7)	31.5	57.8
Weighted sum (λ_1=0.5, λ_2=0.5)	32.9	58.4
Weighted sum (λ_1=0.7, λ_2=0.3)	32.7	58.9
Weighted sum (λ_1=0.9, λ_2=0.1)	32.0	59.3
Ours (λ_1=0.1, λ_2=0.9)	46.7	53.7
Ours (λ_1=0.3, λ_2=0.7)	47.1	54.1
Ours (λ_1=0.5, λ_2=0.5)	48.4	54.6
Ours (λ_1=0.7, λ_2=0.3)	48.9	55.3
Ours (λ_1=0.9, λ_2=0.1)	49.6	55.5

Table A2 Impact of γ.

Method	UCF101-VISPR	
	Action (Top-1 ↑)	Privacy (cMAP (↓)
$\gamma = 0$	45.9	55.7
$\gamma = 0.3$	48.1	54.8
$\gamma = 0.5$	48.4	54.6
$\gamma = 0.7$	47.7	55.2
$\gamma = 1$	43.4	57.2

Table A3 Impact of b.

Method	UCF101-VISPR	
	Action (Top-1 ↑)	Privacy (cMAP (↓)
$b = 0$	43.9	56.9
$b = 0.1$	47.6	55.3
$b = 0.3$	48.4	54.6
$b = 0.5$	48.1	55.1
$b = 1$	47.3	55.4
$b = 2$	47.1	55.9

Appendix B Further Experimental Analysis and Discussion

Analysis of transferability between different utilization tasks. To better investigate the privacy preservation problem with MLLM, we also analyze the generalization ability of our framework between different utilization tasks. Specifically, we train the models on UCF101-VISPR with the action recognition task and evaluate their performance of VQA task on OK-VQA dataset. We report the strongest-performing baseline under the cross-task transfer setting. As shown in Tab. B6, even when our framework is trained on another utilization task (i.e., action recognition), the performance can still be comparable to the performance when the model is trained on the VQA task. Compared with other privacy-preservation methods, our framework also shows lower utility degradation when transferring between tasks, indicating better cross-task generalization ability.

White-box setting. In the experiments in the main paper, we assume that the MLLM is a black-box function with no access to its internal structures and use the estimated gradients to update the anonymizer. We also want to investigate the white-box scenario to better explore this problem, where we relax the constraint and allow gradient backpropagation from the frozen MLLM. Here, we take the VQA task as an example and use the actual gradients from the MLLM to update the anonymizer. The results are shown in Tab. B7. As shown, even in the black-box scenario, our framework with the proposed designs can achieve comparable results with the white-box setting, indicating the efficacy of our proposed designs.

More discussion about downsampling method. In our main experiments, we conduct the downsampling method with downsampling factor of 2. Here, we also investigate the performance of downsampling methods with different downsampling factors. As shown in Tab. B8, when the downsampling factor increases, the performance of the action recognition task drops significantly. Though downsampling methods can protect privacy information in the visual data, they can greatly harm the utility.

Novel action and privacy attributes protocol. In the main paper, we follow [1, 2] to evaluate our framework on the action recognition task with two evaluation settings, in which the action and privacy attributes during training and

Table A4 Impact of m.

Method	UCF101-VISPR		Approximate GPU Memory
	Action (Top-1 ↑)	Privacy (cMAP (↓))	
$m = 0$	35.7	57.5	9760 MB
$m = 5$	46.3	55.3	9795 MB
$m = 10$	48.4	54.6	9868 MB
$m = 20$	48.5	54.6	9994 MB

Table A5 Impact of d.

Method	UCF101-VISPR	
	Action (Top-1 ↑)	Privacy (cMAP (↓))
$d = 0.1$	48.1	54.9
$d = 0.5$	48.4	54.6
$d = 0.9$	46.8	55.7
$d = 1$	46.5	55.8

testing are the same. Here, following [1, 2], we also evaluate our framework on the *novel action and privacy attributes protocol*. In this protocol, for action recognition, the framework is trained using UCF101 [5] training set and subsequently evaluated using samples in PA-HMDB [6]. For privacy, the framework is trained using VISPR [7] training set with 7 privacy attributes (i.e., *gender, complete face, partial face, skin color, seminudity, personal relationship, and social relationship*), denoted as VISPR1. Then, the framework is evaluated using VISPR testing set with 7 different privacy attributes (i.e., *tattoo, race, sports, weight group, age group, hair color, and landmark*), denoted as VISPR2. The action classes and privacy attributes during training and testing do not overlap. We follow [2] to evaluate our framework on this setting. The results are shown in Tab. B9. As can be seen, our framework achieves good generalization performance with novel action and privacy classes.

More experiments on the same-dataset evaluation. To further evaluate our framework, we also conduct experiments on the recent VP-HMDB51 and VP-UCF101 benchmarks [9], which contain video-level annotations of the privacy attributes for the UCF101 dataset and the HMDB51 datasets. Following [9], we adopt ViT-S [11–13] pre-trained on ImageNet [14] as the privacy reviewer f_P, which predicts the privacy attributes on the video level. We follow the settings in [9] and evaluate our framework on VP-HMDB51 and VP-UCF101 benchmarks. The results are shown in Tab. B10. As shown, our framework can achieve better trade-offs on these benchmark, demonstrating its effectiveness.

Per-category privacy analysis. While cMAP provides a concise summary of overall privacy preservation performance, different privacy attributes exhibit varying levels of difficulty. We therefore conduct a per-category analysis of privacy attributes in the Supplementary. Specifically, on UCF101-VISPR, following the privacy attributes defined in VISPR, we evaluate our method on the following privacy categories: gender, complete face, partial face, skin color, seminudity, personal relationship, and social relationship. For each privacy category, we report the per-category precision metric, calculated by $TP_i/(TP_i + FP_i)$. TP_i and FP_i are the numbers of true positives and false positives of privacy attribute category i. Tab. B11 shows the breakdown across privacy categories. Interestingly, the results reveal nuanced patterns: (1) Our method achieves the most substantial privacy reduction for personal and social relationship attributes. This suggests that the proposed black-box MLLM optimization can effectively anonymize relational and contextual cues while preserving the core action semantics required for downstream understanding. (2) On the other hand, gender remains the most difficult attribute to anonymize. This is likely because gender information is implicitly encoded across multiple visual factors (such as body shape, clothing style, and hairstyle) that are also informative for action recognition and scene understanding in MLLMs, making complete disentanglement particularly challenging.

Analysis on time and cost. As shown in Tab. 6 of the main manuscript, our method substantially reduces the number of MLLM calls per sample compared to standard zeroth-order optimization methods. This reduction in MLLM queries directly translates to lower wall-clock time and cost of training. Specifically, in our experiments on the UCF101-VISPR benchmark, training with our method reaches stable performance

Table B6 Results of the transferability from action recognition to VQA task.

Method	**OK-VQA** VQA (Accuracy ↑)	**OK-VQA** Privacy (cMAP ↓)
Baseline [1] trained on action recognition task	39.8	50.8
Baseline [1] trained on VQA task	43.1	48.2
Ours trained on action recognition task	56.1	47.9
Ours trained on VQA task	57.6	44.2

Table B7 Results of white-box setting on the VQA task.

Method	**OK-VQA** VQA (Accuracy ↑)	**OK-VQA** Privacy (cMAP ↓)	**VISPR** Privacy (cMAP ↓)
White-box setting	57.9	44.7	48.9
Black-box setting	57.6	44.2	48.8

in approximately 5 hours under our experimental setup, with an API cost of around $450 when using GPT-4V, representing approximately an 82% reduction compared to training with the baseline SPSA.

Appendix C Further Elaboration and Analysis on Pareto Optimality

C.1 Elaboration on the Weighted Sum Objective

Here, we provide more elaborations on the weighted sum objective and discuss its limitations in handling non-convex problems, as an extension of the discussion in Sec. 3.1 in the main paper.

Specifically, consider two objectives (L_1 and L_2). We provide demonstrations for the trade-off surfaces in convex and non-convex cases as the dotted curves in Fig. C3, where the trade-off surface is the collection of Pareto optimal solutions. In this example, the weighted sum objective becomes $L_{ws}(\theta) = \lambda_1 L_1(\theta) + \lambda_2 L_2(\theta)$. Intuitively, as shown in Fig. C3, the weighted sum objective "draws straight lines" in the problem space and finds where they touch the trade-off surface. When the objectives and the trade-off surface are convex, the weighted sum objective works well. As shown in Fig. C3(a), it can approach all the points on the trade-off surface by adjusting the weights (λ_1 and λ_2) properly.

However, when the trade-off surface is non-convex, as shown in Fig. C3(b), the weighted sum objective cannot find those Pareto optimal solutions in the non-convex regions of the trade-off surface (i.e., the orange dots in Fig. C3 (b)). Yet, it is possible that in the non-convex regions, there exist some solutions that can achieve a relatively smaller sacrifice in one objective while making a significant gain in the other. Such solutions in the non-convex regions could be more advantageous, but is not reachable by the weighted sum objective. Also, because using weighted sum objective cannot reach the Pareto optimal solutions in the non-convex part, it may not be able to generate diverse solutions with different trade-offs. This can be undesirable for real-world problems, where diverse solutions are needed to offer different types of trade-offs for different cases.

In summary, weighted sum objective is not suitable for our problem. This is also consistent with the experiment results where the weighted sum objective does not achieve satisfactory performances (see Tab. 5 in the main paper and also Tab. A1 in this Supplementary).

C.2 Elaboration on the Proof of Theorem 3.1

In our method, we design the overall objective to optimize the anonymizer f_A with the augmented Tchebycheff objective (Eq. 4 in the main paper), which has favorable properties of allowing us to achieve an optimal trade-off between the privacy and the black-box MLLM's utility, as analyzed in Theorem 3.1. Here we further provide elaborations on the proof for Theorem 3.1 in the main paper.

Table B8 More results with downsampling method [2].

Method	HMDB51-VISPR Action Top-1 (↑)	HMDB51-VISPR Privacy cMAP (↓)	UCF101-VISPR Action Top-1 (↑)	UCF101-VISPR Privacy cMAP (↓)
Downsample-2×	42.1	61.2	43.1	57.2
Downsample-4×	33.9	41.4	39.5	50.1
Downsample-8×	23.5	33.7	27.5	43.1

Table B9 Results on novel action and privacy attributes protocol. VISPR1 denotes the training samples with 7 privacy attributes. VISPR2 denotes the testing samples with 7 different privacy attributes. Details about the privacy attributes are introduced above.

Method	UCF101→PA-HMDB Action Top-1 (↑)	VISPR1→VISPR2 Privacy cMAP (↓)
Raw data	50.3	57.6
Downsample [2]	42.1	52.2
Obf-Blackening [2]	26.6	53.6
Obf-StrongBlur [2]	27.2	53.7
Obf-WeakBlur [2]	33.2	55.8
Surrogate model of VITA [6]	30.0	52.8
Surrogate model of SPAct [2]	30.2	50.3
Surrogate model of BDQ [8]	32.1	50.5
Surrogate model of MPPAR [1]	35.8	49.8
Surrogate model of STPrivacy [9]	34.6	50.4
SelectivePrivacy [10]	37.9	55.8
Ours	**45.2**	**48.8**

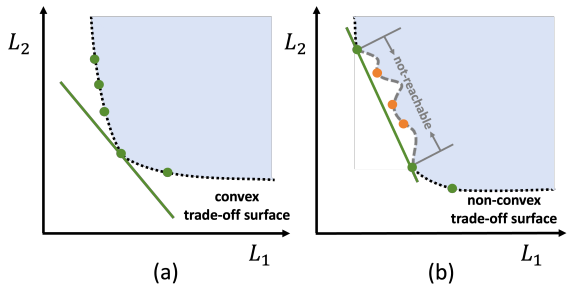

Fig. C3 Illustration of the weighted sum objective in convex and non-convex scenarios. The dotted curves represent the trade-off surface, and the green lines indicate the weighted-sum objective. The green dots on the trade-off surfaces represents Pareto optimal solutions that can be achieved by the weighted sum objectives by adjusting weights, while the orange dots on the trade-off surface in (b) represents the Pareto optimal solutions that cannot be reached by the weighted sum objective.

For convenience, we repeat the augmented Tchebycheff objective (Eq. 4 in the main paper) here:

$$\min_{\theta} \; L_{AT}(\theta) = \min_{\theta} \Big(\max_{i} \; \{\lambda_i \big(L_{T_i}(\theta) - (z_i^* - \varepsilon)\big)\} + \sigma \sum_{i} \lambda_i L_{T_i}(\theta) \Big) \tag{C1}$$

where z_i^* is the ideal value of each objective, ε is a small positive scalar, and σ is a sufficiently small positive scalar.

We also repeat the definitions of Pareto dominance, Pareto optimality, and Theorem 3.1 here:

Definition 1 (Pareto Dominance) Consider n objectives, and let $\theta_1, \theta_2 \in \mathbb{R}^{p_A}$. θ_1 is said to dominate θ_2, if and only if $L_i(\theta_1) \leq L_i(\theta_2)$, $\forall i \in \{1, ..., n\}$, and besides $L_j(\theta_1) < L_j(\theta_2)$, $\exists j \in \{1, ..., n\}$.

Definition 2 (Pareto Optimality) A solution $\theta^* \in \mathbb{R}^{p_A}$ is Pareto optimal, if there exists no θ that dominates θ^*.

Theorem 1 *Let $\theta \in \mathbb{R}^{p_A}$. θ is Pareto optimal if and only if θ solves Eq. C1 with $\lambda = [\lambda_1, \lambda_2] \in \mathbb{R}^2$, $\lambda_1, \lambda_2 > 0$.*

Below, we first show that if θ solves Eq. C1, it is Pareto optimal.

Suppose $\theta_p \in \mathbb{R}^{p_A}$ solves (minimizes) Eq. C1 and θ_p is not Pareto optimal. Then there exists $\theta_m \in \mathbb{R}^{p_A}$ that dominates θ_p, i.e., $L_{T_i}(\theta_m) \leq L_{T_i}(\theta_p)$, $\forall i \in \{1, 2\}$ and $L_{T_j}(\theta_m) < L_{T_j}(\theta_p)$, $\exists j \in \{1, 2\}$. Thus, we have $\lambda_i \big(L_{T_i}(\theta_m) -$

Table B10 Results on VP-HMDB51 and VP-UCF101.

Method	VP-HMDB51 Action Top-1 (↑)	VP-HMDB51 Privacy cMAP (↓)	VP-UCF101 Action Top-1 (↑)	VP-UCF101 Privacy cMAP (↓)
Raw data	45.4	75.6	55.4	76.6
Downsample [2]	36.6	71.4	43.1	71.5
Obf-Blackening [2]	29.8	74.1	40.1	75.4
Obf-StrongBlur [2]	31.0	74.3	43.6	75.6
Obf-WeakBlur [2]	36.7	75.1	46.0	76.0
Surrogate model of VITA [6]	31.2	74.0	35.2	74.8
Surrogate model of SPAct [2]	31.5	73.6	34.4	74.2
Surrogate model of BDQ [8]	32.9	73.4	34.8	73.2
Surrogate model of MPPAR [1]	34.8	72.9	36.0	73.7
Surrogate model of STPrivacy [9]	35.8	71.4	38.9	72.1
SelectivePrivacy [10]	36.8	73.8	42.5	74.9
Ours	**41.4**	**69.9**	**52.3**	**69.7**

Table B11 Per-category privacy preservation performance.

	Per-category precision							Overall (cMAP)
	Gender	Complete face	Partial face	Skin color	Semi-nudity	Personal relationship	Social relationship	
Ours	69.2	59.5	59.2	57.2	58.6	40.2	38.2	54.6

$(z_i^* - \varepsilon)\big) \leq \lambda_i\big(L_{T_i}(\theta_p) - (z_i^* - \varepsilon)\big)$, $\forall i \in \{1, 2\}$ and $\sigma \sum_i \lambda_i\big(L_{T_i}(\theta_m)\big) < \sigma \sum_i \lambda_i\big(L_{T_i}(\theta_p)\big)$. This then means that $L_{AT}(\theta_m)$ has a smaller value than $L_{AT}(\theta_p)$, which contradicts with the assumption that θ_p minimizes Eq. C1. Thus, there does not exist θ_m such that θ_m dominates θ_p. Then, by the definition of Pareto optimality, θ_p is Pareto optimal.

Then, we show that θ is Pareto optimal implies θ solves augmented Tchebycheff objective for some λ.

Let $\theta_p \in \mathbb{R}^{p_A}$ be Pareto optimal, i.e., there exists no $\theta_n \in \mathbb{R}^{p_A}$ that dominates θ_p. Thus, for some $\lambda_1, \lambda_2 > 0$, we have $\max_i \{\lambda_i\big(L_{T_i}(\theta_p) - (z_i^* - \varepsilon)\big)\} < \max_i \{\lambda_i\big(L_{T_i}(\theta_n) - (z_i^* - \varepsilon)\big)\}$ [15].

Suppose there exist $\theta_m \in \mathbb{R}^{p_A}$ which solves Eq. C1 with λ and $\theta_m \neq \theta_p$. Then the lower bound of Eq. C1 is given by $L_{AT}(\theta_m) = \max_i \{\lambda_i\big(L_{T_i}(\theta_m) - (z_i^* - \varepsilon)\big)\} + \sigma \sum_i \lambda_i L_{T_i}(\theta_m)$. Now, θ_p remains minimality of Eq. C1 if $L_{AT}(\theta_p) < L_{AT}(\theta_m)$, i.e.:

$$\begin{aligned} &\sigma \sum_i \lambda_i\big(L_{T_i}(\theta_p) - L_{T_i}(\theta_m)\big) \\ &< \max_i \{\lambda_i\big(L_{T_i}(\theta_m) - (z_i^* - \varepsilon)\big)\} \\ &\quad - \max_i \{\lambda_i\big(L_{T_i}(\theta_p) - (z_i^* - \varepsilon)\big)\} \end{aligned} \tag{C2}$$

This means that for all $\theta_m \in \mathbb{R}^{p_A}$ with $\sum_i \lambda_i\big(L_{T_i}(\theta_p) - L_{T_i}(\theta_m)\big) > 0$, to ensure θ_p minimizes Eq. C1, we need to have:

$$\sigma < \frac{\max_i \{\lambda_i\big(L_{T_i}(\theta_m) - (z_i^* - \varepsilon)\big)\} - \max_i \{\lambda_i\big(L_{T_i}(\theta_p) - (z_i^* - \varepsilon)\big)\}}{\sum_i \lambda_i\big(L_{T_i}(\theta_p) - L_{T_i}(\theta_m)\big)} \tag{C3}$$

Thus, for σ defined as a sufficiently small positive scalar in Eq. C1, θ_p holds minimality of Eq. C1, i.e., θ_p solves Eq. C1 with λ.

Overall, we elaborate on (1) if θ solves Eq. C1, then θ is Pareto optimal; and (2) if θ is Pareto optimal, then θ is also a solution to Eq. C1. Thus we have Theorem 1 (i.e., Theorem 1 in the main paper).

Appendix D Algorithm for Critical-History Enhanced Optimization

In Sec. 3.2 of the main paper, we introduce our proposed critical-history enhanced optimization. Algorithm 1 shows the detailed steps for this mechanism. We show how to obtain the enhanced gradient $\tilde{g}_k$ and status indicator I_k using history information in lines 1-4, and how to construct the critical gradient collection set in lines 6-19.

Algorithm 1 Zeroth-Order Gradient Estimation with Critical History Gradient

Require: Input user visual data X, black-box MLLM, anonymizer f_A with learnable parameters $\theta_k \in \mathbb{R}^{p_A}$, positive scalars c_k, γ, b, d, history gradient collection set C_g with capacity m, optimization step k

Generate the perturbation vector $\Delta_k \in \mathbb{R}^{p_A}$ with each of its element randomly sampled from Bernoulli ± 1 distribution

$\hat{g}_k(\theta_k) \leftarrow \frac{\left(L_{MLLM}(\theta_k + c_k\Delta_k; X) - L_{MLLM}(\theta_k - c_k\Delta_k; X)\right)}{2c_k\Delta_k}$ ▷ Eq. 5 in main paper

$\tilde{g}_k \leftarrow \gamma\hat{g}_k + \frac{1-\gamma}{m}\sum C_g$ ▷ Eq. 7 in main paper

$I_k \leftarrow \frac{b}{\pi}\arctan(2\pi\cos\varphi_k)$ ▷ Eq. 8 in main paper

$C_g \leftarrow \text{SelectCriticalGradients}(\hat{g}_k, C_g, d)$

function SelectCriticalGradients($\hat{g}_k$, C_g, d)

  Initialize S as an empty set

  $s_k \leftarrow \|\hat{g}_k\|_2$ ▷ Eq. 6 in main paper

  for $\hat{g}_{t_i}$ in C_g **do**

    $s_{t_i} \leftarrow d^{k-t_i}\|\hat{g}_{t_i}\|_2$ ▷ Eq. 6 in main paper

    S.add(s_{t_i})

  end for

  if $s_k > s_{t_j},\ s_{t_j} = \min S$ **then**

    C_g.add($\hat{g}_k$)

    C_g.remove($\hat{g}_{t_j}$)

  else if len(C_g) $< m$ **then**

    C_g.add($\hat{g}_k$)

  end if

  return C_g

end function

Appendix E More Experiment Details

More details about evaluation metrics. In our experiments, we follow [1, 2, 6, 9, 16] to adopt cMAP metric to evaluate the performance of privacy preservation. More specifically, the performance of privacy preservation is evaluated by using a privacy attribute classifer to predict the privacy information (e.g., face, skin color, etc.) from the anonymized data. The accuracy of this prediction is measured via classwise mean average precision (cMAP) metric, which can reflect the performance of privacy preservation: lower accuracy of this prediction (i.e., lower cMAP) indicates lower privacy leakage in the anonymized data and better privacy preservation. The cMAP metric measures how accurately the privacy attribute classifier can predict privacy information from the anonymized data, which reflects the performance of privacy preservation. Lower value in cMAP means it is harder to recognize privacy information in the anonymized data, i.e., lower value in cMAP indicates lower privacy leakage. Specifically, cMAP is calculated by:

$$cMAP = \frac{1}{n}\sum_{i=1}^{n}(TP_i/(TP_i + FP_i))$$

where TP_i and FP_i are the numbers of true possitives and false possitives of privacy attribute class i, and n is the total number of privacy attribute classes.

More details about evaluation benchmarks and annotations. In the experiments in the main paper, we evaluated our framework on privacy-preserving action recognition benchmarks, including HMDB51-VISPR [2, 6] and UCF101-VISPR [2]. Specifically, following previous works [1, 2], the definition of privacy (i.e., privacy attributes) in are as follows: for HMDB51-VISPR, the privacy attributes are *gender, complete face, partial face, skin color, semi-nudity, and personal relationship*; for UCF101-VISPR, the privacy attributes are *gender, complete face, partial face, skin color, semi-nudity, personal relationship, and social relationship*. For the action recognition experiments, we follow the retrieval-based evaluation approach in [17] to obtain the action classes from the open-ended MLLM responses. The action class with the highest retrieval score is regarded as the top-1 action class. For the VQA experiments, we follow [18–20] to perform and evaluate the VQA task with MLLM.

In our experiments with the VQA task, we annotate the testing images with the same privacy attribute labels following the definition of privacy in previous works [2] (i.e., *gender, complete face, partial face, skin color, semi-nudity, personal relationship, and social relationship*). We asked 3 annotators to review each image and assign a binary label for each privacy attribute.

The final labels are determined by majority voting of the annotations following previous work [9]. The annotations are available at link.

More details about f_A, f_P, L_{MLLM}, and $L_{Privacy}$. Following previous works [1, 2, 6], we build our anonymizer f_A based on UNet [21]. To reduce the number of learnable parameters as discussed in [22, 23] to effectively perform gradient estimation, we make small modifications to the UNet. Specifically, we reduce the feature channels of the intermediate features in the UNet following [24] and adopt the lightweight depth-aware separable convolution layers following [25] to process the features and generate the final output. Then, following [2], we initialize the anonymizer by training it with image reconstruction loss. After that, we only update the 9k parameters of the lightweight convolution layers (instead of 25M parameters of the original UNet) using Eq. 9 in the main paper.

Following [2, 6], we adopt ResNet-50 [26] classifier as the privacy reviewer. We follow the same training strategy as previous methods [1, 6, 16] for the privacy reviewer.

In action recognition experiments, during training, we obtain the MLLM utilization objective L_{MLLM} using the cross-entropy between the ground truth and the scores of each action class obtained in the retrieval-based MLLM evaluation approach [17]. In VQA experiments, following [27, 28], to obtain L_{MLLM}, we use CLIP cosine distance score [29] between the MLLM output and the ground truth. For both tasks, the privacy objective during training is obtained following [6].

More details about the Augmented Tchebycheff objective and other hyperparameters We set z_i^* in Eq. 4 in the main paper to be the current best (smallest) value of each objective, set ε to be $0.1 \times z_i^*$, and set $\sigma = 0.0001$. This setting is shown to be effective in all our experiments. Following standard SPSA criteria [22, 23], we formulate the step size factor $\{a_k\}$ with initial value 0.01 and exponential decay rate 0.4, and formulate the factor $\{c_k\}$ with initial value of 0.005 and exponential decay rate 0.2. We set the weights in Eq. 4 in the main paper to be $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$. The decay factor in Eq. 6 in the main paper is set as $d = 0.5$. The factor controlling the impact of history gradient in Eq. 7 in the main paper is set as $\gamma = 0.5$. The factor controlling the magnitude of status indicator I_k in Eq. 8 of the main paper is set to be $b = 0.3$. We construct the critical gradient collection set with capacity $m = 10$. Discussions about the hyperparameters are provided in Sec. A.

More details about comparison methods. We follow [2] to implement the *Downsample* and *Obfuscation-based* methods in Tab. 1 and Tab. 2 (main paper). For the previous privacy-preserving action recognition methods [1, 2, 6, 8, 9], we follow [30] and consider the action recognition model in their frameworks as a surrogate for the black-box MLLM for the action recognition task. We use the gradients backpropagated from the action recognition model to update the anonymizer. For the VQA task, we follow the same pipeline and replace the action recognition model with [31].

Appendix F Licenses

Dataset licenses. We use VISPR dataset [7] following Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License. We use HMDB51 dataset [32], UCF101 dataset [5], and OK-VQA dataset [18] following Creative Commons Attribution 4.0 International (CC BY 4.0) License.

MLLM licenses. We use Video-LLaVA [33] and LLaVA [19] following Apache License 2.0. We use GPT-4V [34] following the terms of using OpenAI services.

References

[1] Peng, D., Xu, L., Ke, Q., Hu, P., Liu, J.: Joint attribute and model generalization learning for privacy-preserving action recognition. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)

[2] Dave, I.R., Chen, C., Shah, M.: Spact: Self-supervised privacy preservation for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20164–20173 (2022)

[3] Paul, M., Ganguli, S., Dziugaite, G.K.: Deep learning on a data diet: Finding important examples early in training. Advances in Neural Information Processing Systems **34**, 20596–20607 (2021)

[4] McRae, P.-A.M., Parthasarathi, P., Assran, M., Chandar, S.: Memory augmented optimizers for deep learning. In: International Conference on Learning Representations (2022). https://openreview.net/forum?id=NRX9QZ6yqt

[5] Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)

[6] Wu, Z., Wang, H., Wang, Z., Jin, H., Wang, Z.: 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 (2020)

[7] Orekondy, T., Schiele, B., Fritz, M.: Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)

[8] Kumawat, S., Nagahara, H.: Privacy-preserving action recognition via motion difference quantization. In: European Conference on Computer Vision, pp. 518–534 (2022). Springer

[9] Li, M., Liu, J., Fan, H., Liu, J.-W., Li, J., Shou, M.Z., Keppo, J.: Stprivacy: Spatio-temporal tubelet sparsification and anonymization for privacy-preserving action recognition. arXiv preprint arXiv:2301.03046 (2023)

[10] Ilic, F., Zhao, H., Pock, T., Wildes, R.P.: Selective interpretable and motion consistent privacy attribute obfuscation for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18730–18739 (2024)

[11] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., *et al.*: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)

[12] Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.-J.: Dynamicvit: Efficient vision transformers with dynamic token sparsification. Advances in neural information processing systems **34**, 13937–13949 (2021)

[13] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357 (2021). PMLR

[14] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848

[15] Miettinen, K.: Nonlinear Multiobjective Optimization. International Series in Operations Research & Management Science. Springer, ??? (1999). https://books.google.com.sg/books?id=ha_zLdNtXSMC

[16] Wu, Z., Wang, Z., Wang, Z., Jin, H.: Towards privacy-preserving visual recognition via adversarial training: A pilot study. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 606–624 (2018)

[17] Li, S., Zhang, Y., Zhao, Y., Wang, Q., Jia, F., Liu, Y., Wang, T.: Vlm-eval: A general evaluation on video large language models. arXiv preprint arXiv:2311.11865 (2023)

[18] Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

[19] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)

[20] Chen, J., Zhu, D., Shen, X., Li, X., Liu,

Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478 (2023)

[21] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234–241 (2015). Springer

[22] Spall, J.C.: Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. IEEE transactions on automatic control **37**(3), 332–341 (1992)

[23] Oh, C., Hwang, H., Lee, H.-y., Lim, Y., Jung, G., Jung, J., Choi, H., Song, K.: Blackvip: Black-box visual prompting for robust transfer learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 24224–24235 (2023)

[24] Ma, Z., Chang, D., Xie, J., Ding, Y., Wen, S., Li, X., Si, Z., Guo, J.: Fine-grained vehicle classification with channel max pooling modified cnns. IEEE Transactions on Vehicular Technology **68**(4), 3224–3233 (2019) https://doi.org/10.1109/TVT.2019.2899972

[25] Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)

[26] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[27] Ouali, Y., Bulat, A., Martinez, B., Tzimiropoulos, G.: Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024, pp. 395–413. Springer, Cham (2025)

[28] Chen, Z., Zhou, Q., Shen, Y., Hong, Y., Sun, Z., Gutfreund, D., Gan, C.: Visual chain-of-thought prompting for knowledge-based visual reasoning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 1254–1262 (2024)

[29] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PMLR

[30] Liu, Y., Chen, X., Liu, C., Song, D.: Delving into transferable adversarial examples and black-box attacks. arXiv preprint arXiv:1611.02770 (2016)

[31] Gardères, F., Ziaeefard, M., Abeloos, B., Lecue, F.: Conceptbert: Concept-aware representation for visual question answering. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 489–498 (2020)

[32] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: 2011 International Conference on Computer Vision, pp. 2556–2563 (2011). IEEE

[33] Lin, B., Zhu, B., Ye, Y., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122 (2023)

[34] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)