Boundary Probing for Input Privacy Protection When Using LMM Services

Supplementary Material

A. More Qualitative Results and Ablation Studies

A.1. More Qualitative Results

We provide qualitative results of the anonymized data generated using our PABP framework for VQA in Fig. 1 and more qualitative results for action recognition in Fig. 2. We also show more visualizations of the anonymized data generated by our method on the testing set of VISPR dataset [28] in Fig. 3. As can be observed, our method effectively maintains the LMM's performance and can successfully remove private information (e.g., face, nudity, and credit card).

A.2. More Ablation Studies

Impact of the decision boundary. Our framework explores the decision boundary between "satisfactory" and "unsatisfactory" LMM utility states to guide the optimization with access only to the LMM's final label outputs. Alternatively, as mentioned in Sec. 3.1 in the main paper, by assigning "satisfactory" LMM utility as 1 and "unsatisfactory" as -1 as the LMM utility loss function, it is also possible to estimate gradients for the black-box LMM to update the anonymization model. Here we investigate this approach. Specifically, we conduct experiments with the following baselines (using the above defined LMM utility loss function) and compare them with our method. In *Random search*, we adopt the random search method [2]. In *Reinforcement learning*, we employ reinforcement learning mechanism [16]. In *Zeroth-order optimization*, we employ zeroth-order optimization [22] to estimate gradients for the black-box LMM's utility. As shown in Tab. 1, our method significantly outperforms all baselines, demonstrating its effectiveness.

Method	Action (Acc.↑)	Privacy (cMAP↓)
Random Search [2]	30.1	62.9
Reinforcement Learning [16]	32.6	61.3
Zeroth-order Optimization [22]	35.5	60.1
Ours	47.9	54.3

Table 1. Impact of decision boundary.

Impact of the utility budget. In our PABP framework, the LMM's utility status is determined with a utility budget τ. Here, we investigate the impact of different utility budgets. As shown in Tab. 2, increasing τ leads to better LMM's utility, while decreasing τ shows stronger privacy protection performance. Thus, by controlling τ, we can obtain different trade-offs between privacy and utility.

Method	Action (Acc.↑)	Privacy (cMAP↓)
$\tau = 0.8$	42.0	50.1
$\tau = 0.85$	45.3	52.7
$\tau = 0.9$	47.9	54.3
$\tau = 0.95$	50.0	57.4

Table 2. Impact of utility budget.

More experiments on the GEP scheme. In our GEP scheme, we first initialize the anonymization model and then apply probing scheme to update it to the "satisfactory" side. In Tab. 5 of the main paper, to evaluate this design, we have compared with the variant *without probing* that skips the probing in GEP. We here further investigate the impact of initializing the model with different surrogate utility models (VGG [32], R3D [9], and ViT [7]). As shown in Tab. 3, using different initialization, the performances of our framework remain stable and all outperform previous methods in Tab. 1 in the main paper, showing the effectiveness of our design.

Method	Action (Acc.↑)	Privacy (cMAP↓)
Initialization with VGG	47.7	54.3
Initialization with R3D	47.9	54.3
Initialization with ViT	47.8	54.1

Table 3. Impact of different initialization. Note that our framework with different initializations all outperform previous methods.

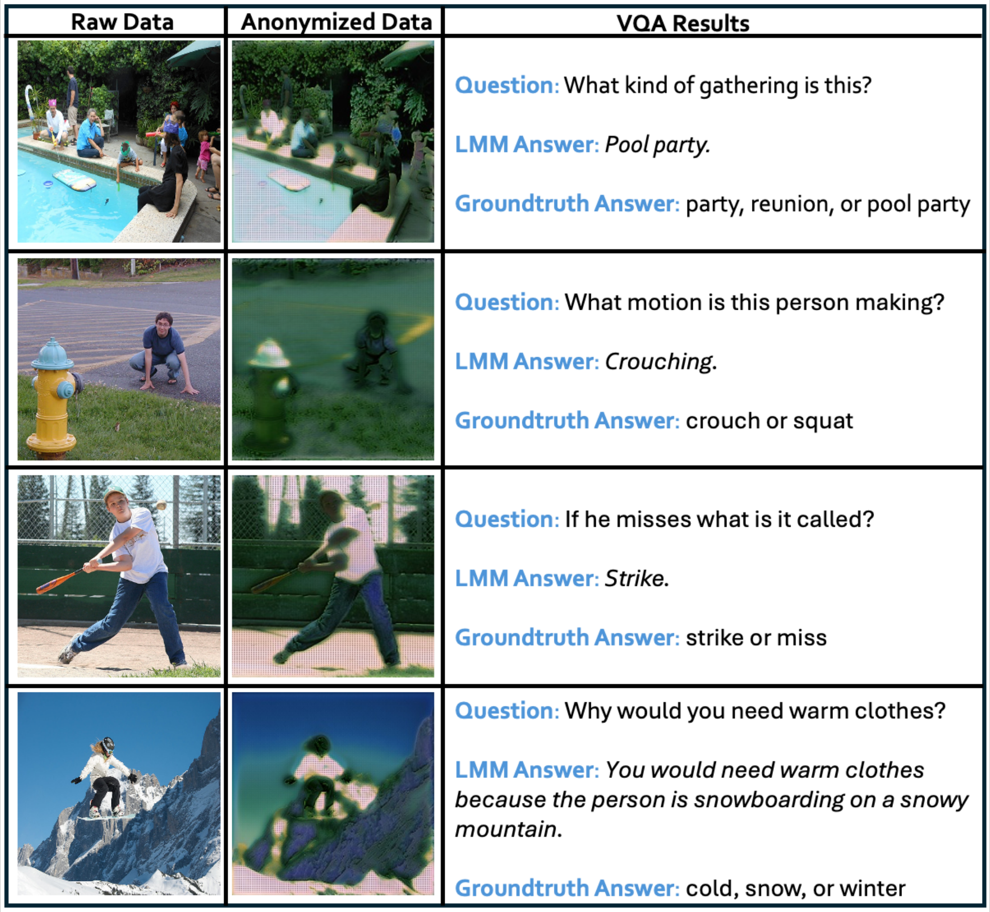
Impact of the radius d in the GEP scheme. In our GEP scheme, we start probing the decision boundary with a sphere with radius d. Here we investigate its impact on the performance of our PABP framework. As shown in Tab 4, our PABP stably achieves good performance with different d. We adopt $d = 0.05$ in our main experiment.

Method	Action (Acc.↑)	Privacy (cMAP↓)
$d = 0.005$	47.5	54.4
$d = 0.05$	47.9	54.3
$d = 0.5$	47.6	54.5
$d = 5$	47.8	54.6

Table 4. Impact of initial radius d.

Impact of Hessian Approximation. In the PGP scheme, we follow the common and efficient practice [20, 33] to use the Fisher Information Matrix (FIM) to approximate diagonal Hessian. Here we also approximate Hessian via other approaches (e.g., L-BFGS [18] and K-FAC [24]) to derive loss contour line and conduct the PGP scheme. As shown in Tab. 5, our framework consistently achieves good performance with different methods to approximate Hessian.

Training time. We show the approximated training time of our PABP framework in Tab. 6.



Question: What kind of gathering is this?

LMM Answer: *Pool party.*

Groundtruth Answer: party, reunion, or pool party

Question: What motion is this person making?

LMM Answer: *Crouching.*

Groundtruth Answer: crouch or squat

Question: If he misses what is it called?

LMM Answer: *Strike.*

Groundtruth Answer: strike or miss

Question: Why would you need warm clothes?

LMM Answer: *You would need warm clothes because the person is snowboarding on a snowy mountain.*

Groundtruth Answer: cold, snow, or winter

Figure 1. Qualitative results of anonymized VQA images. For each row, we show the raw image, the anonymized image generated using our method, the question, the LMM answer, and the groundtruth answer. The LMM answer is generated by feeding the fixed off-the-shelf LMM (LLaVA) with the anonymized image and the question. An LMM answer is considered to be correct if it matches the groundtruth answer following the evaluation rule in [1, 3], and the LMM answers in the examples shown above are all considered correct.

Method	Action (Acc.↑)	Privacy (cMAP↓)
PABP with L-BFGS	47.9	54.7
PABP with K-FAC	47.6	54.4
PABP with diagonal FIM	47.9	54.3

Table 5. Results of using different methods to approximate Hessian.

Approximated Training Time	
GEP	2 hrs
PGP	16 hrs
Total training time	18 hrs

Table 6. Training time.

B. Further Analysis

White-box setting. In the experiments in the main paper, we consider the LMM as a black-box model where we only have access to its output. Here, as an investigation, we also explore the white-box scenario where we relax the constraint and allow gradient backpropagation from the frozen LMM. We take the VQA task as an example to use the obtained gradient from LMM (LLaVA) to update the anonymization model. As shown in Tab. 7, even with the black-box scenario, our PABP can achieve results close to the white-box setting, showing its efficacy.

Method	VQA (Acc.↑)	Privacy (cMAP↓)	VISPR (cMAP↓)
White-box setting	58.5	44.0	48.9
Black-box setting	57.3	44.2	51.4

Table 7. Results of white-box setting.

Analysis on the Transferability between utility tasks. We investigate the transferability of our method between different utility tasks. Specifically, during training, we train the anonymization model w.r.t. the action recognition task on UCF101-VISPR benchmark [35], and evaluate on the VQA task. As shown in Tab. 8, even directly applying the trained anonymization model with another utility task (i.e., action recognition), the LMM's performance on the

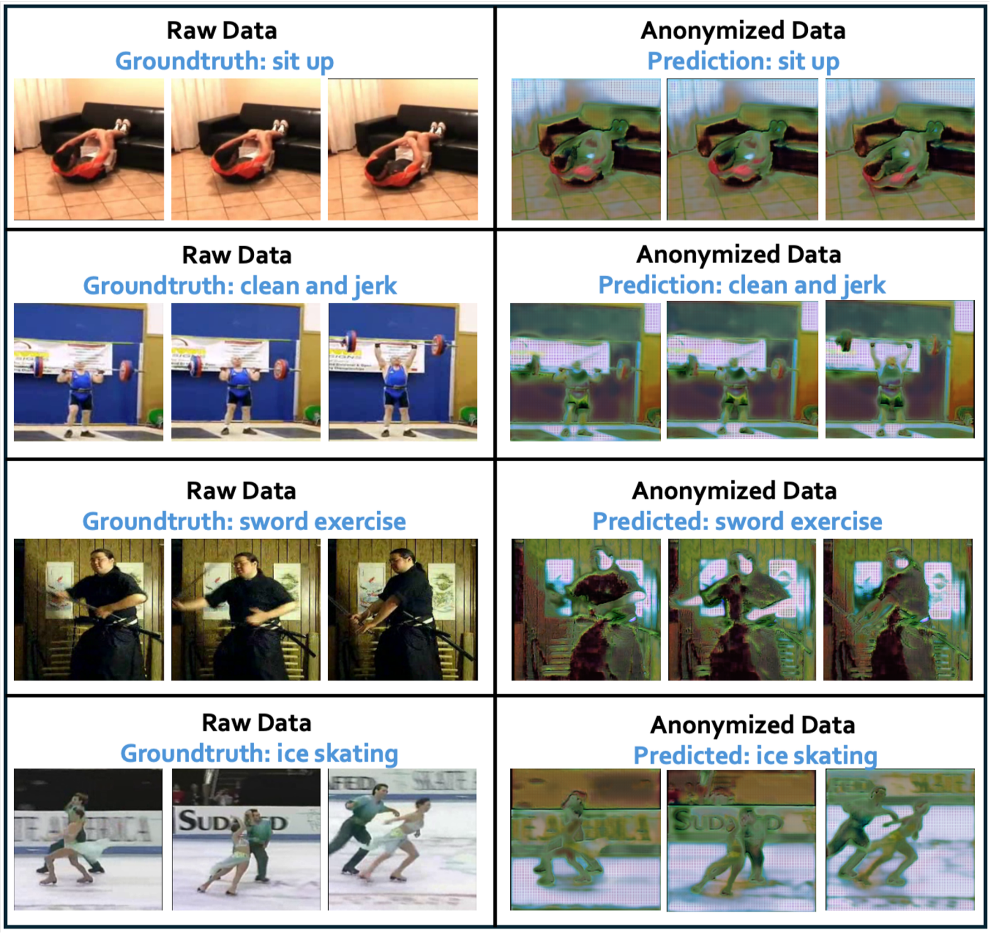


Figure 2. Qualitative results of anonymized data (action video frames). In each row, we show the raw data, the groundtruth action class, the anonymized data, and the predicted action class. As shown, the LMM (Video-LLaVA) can correctly predict the action class given the anonymized data, while the privacy information such as faces and skin color in the anonymized data is protected.

VQA task can still be comparable to the performance when training on the VQA task, showing the transferability of our framework across utility task.

Method	VQA (Acc.↑)	Privacy (cMAP↓)
Trained with action recognition	56.2	45.7
Trained with VQA	57.3	44.2

Table 8. Results of transferability between utility tasks.

More analysis about the downsampling method. In our main experiments, we conduct the downsampling method [6] with downsampling factor of 2. Here we also investigate the impact of different downsampling factor values. As shown in Tab. 9, when increasing the downsampling factor, the performance of action recognition drops significantly. Though downsampling can protect privacy information, they can greatly degrade the utility.

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

Table 9. More results with the downsampling method [6].

C. More Details and Analysis about the PGP scheme

We introduce the PGP scheme in Sec. 3.1 in the main paper. Specifically, when updating the anonymization model ϕ_k (at the k-th step) using the gradient descent point (i.e., $\phi'_k = \phi_k - \alpha g_k$) will bring the model to the "unsatisfactory" side of the decision boundary, the PGP scheme facilitates to maintain the update within the "satisfactory" side while achieving the same progress in optimizing privacy protection as gradient descent. Below we provide more details and analysis about the PGP scheme.

More details in the derivation of the contour line. As

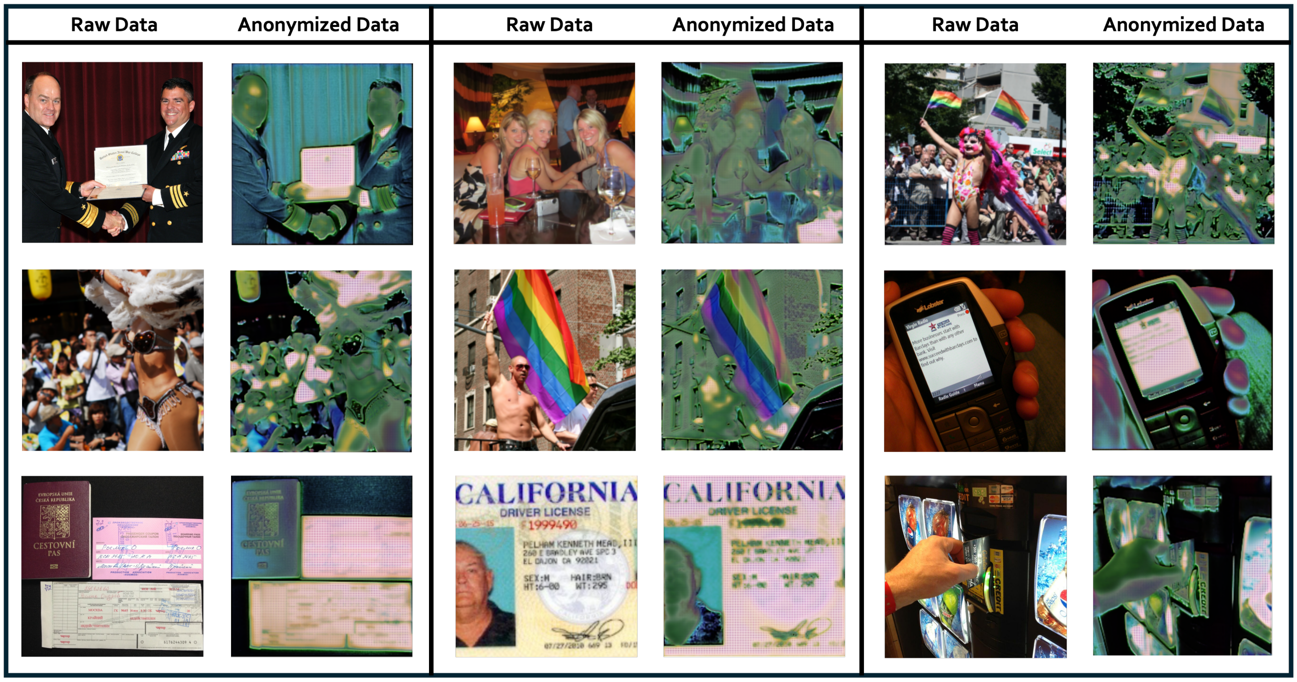


Figure 3. More visualizations of the anonymized data generated by our method on VISPR testing set. Note that, VISPR dataset contains annotations of various privacy attributes such as face, nudity, and credit card. As shown in the figure, our method can effectively remove the expected privacy information.

elaborated in Sec. 3.1 in the main paper, the PGP scheme searches for an alternative point (i.e., $\phi_k + \delta_i$) in the parameter space that satisfies the condition: the point has the same privacy loss value as the gradient descent point (i.e., $L_p(\phi_k + \delta_i) = L_p(\phi'_k)$). To achieve this, PGP solves the contour line of the approximated privacy loss function to sample candidate points over it. Here we provide more details about the derivation of the contour line.

Specifically, based on Taylor expansion, we can approximate the privacy loss $L_p(\phi_k + \delta_i)$ as $m(\delta_i)$ (Eq. 2 in the main paper):

$$m(\delta_i) = L_p(\phi_k) + {\delta_i}^\top g_k + \frac{1}{2}{\delta_i}^\top H_k \delta_i,\ \delta_i \in \Omega, \tag{1}$$

where $\Omega = \{\delta_i \mid \|\delta_i\| \le \varepsilon\}$. Then, we can solve $m(\delta_i) = L_p(\phi')$ to identify points that meet the above condition. In particular, due to the extremely costly computation of the exact Hessian [5, 25], we follow the common and efficient practice [20, 33] to approximate the diagonal Hessian with Fisher Information Matrix. Denoting the j-th diagonal element in H_k as h_k^j, and the j-th element in δ_i as δ_i^j, i.e., $H_k = diag(h_k^1, h_k^2, \dots, h_k^p)$ and $\delta_i = [\delta_i^1, \delta_i^2, \dots, \delta_i^p]^\top$. Then, we have ${\delta_i}^\top H_k \delta_i = \sum_{j=1}^p h_k^j (\delta_i^j)^2$ and can re-write Eq. 1 as $m(\delta_i) = L_p(\phi_k) + \sum_{j=1}^p g_k^j \delta_i^j + \frac{1}{2}\sum_{j=1}^p h_k^j (\delta_i^j)^2$. Then, by setting $m(\delta_i) = L_p(\phi')$, we have Eq. 2 below (i.e., Eq. 3 in the main paper):

$$\sum_{j=1}^p g_k^j \delta_i^j + \frac{1}{2}\sum_{j=1}^p h_k^j (\delta_i^j)^2 = C, \tag{2}$$

where $C = L_p(\phi'_k) - L_p(\phi_k)$.

For cases where $h_k^j \neq 0$ for each j. Specifically, when $h_k^j \neq 0$ for all $j \in [1, 2, \dots, p]$, we can re-write Eq. 2 as:

$$\begin{aligned} &\sum_{j=1}^p \left(g_k^j \delta_i^j + \frac{1}{2} h_k^j (\delta_i^j)^2\right) \\ =&\sum_{j=1}^p \frac{h_k^j}{2}\left((\delta_i^j)^2 + \frac{2g_k^j}{h_k^j}\delta_i^j\right) \\ =&\sum_{j=1}^p \frac{h_k^j}{2}\left((\delta_i^j + \frac{g_k^j}{h_k^j})^2 - (\frac{g_k^j}{h_k^j})^2\right) = C. \end{aligned} \tag{3}$$

We can then organize the above Eq. 3 as Eq. 4 (i.e., Eq. 4 in the main paper) below:

$$\sum_{j=1}^p h_k^j (\delta_i^j + \frac{g_k^j}{h_k^j})^2 - 2C - \sum_{j=1}^p \frac{(g_k^j)^2}{h_k^j} = 0. \tag{4}$$

Eq. 4 describes an p-dimensional ellipsoid centered at $\left(-\frac{g_k^1}{h_k^1}, -\frac{g_k^2}{h_k^2}, \dots, -\frac{g_k^p}{h_k^p}\right)$.

For cases where there are $h_k^j = 0$. When there exists j such that $h_k^j = 0$ in the Hessian, to solve Eq. 2, we can first split the entries of H_k into zeros (i.e., $M = \{j \mid h_k^j = 0\}$) and non-zeros (i.e., $N = \{j \mid h_k^j \neq 0\}$). Note that, $M \cup N = \{1, 2, \dots, p\}$ and $M \cap N = \emptyset$. Then, Eq. 2 can be organized as:

$$\begin{aligned}
&\sum_{j=1}^{p} g_k^j \delta_i^j + \frac{1}{2}\sum_{j=1}^{p} h_k^j (\delta_i^j)^2 \\
&= \sum_{j\in M} g_k^j \delta_i^j + \sum_{j\in N} g_k^j \delta_i^j + \frac{1}{2}\sum_{j\in M} h_k^j (\delta_i^j)^2 + \frac{1}{2}\sum_{j\in N} h_k^j (\delta_i^j)^2 \\
&= \sum_{j\in M} g_k^j \delta_i^j + \sum_{j\in N} \left(g_k^j \delta_i^j + \frac{1}{2} h_k^j (\delta_i^j)^2\right) = C.
\end{aligned} \tag{5}$$

As $\sum_{j\in N}\left(g_k^j\delta_i^j + \frac{1}{2}h_k^j(\delta_i^j)^2\right) = \sum_{j\in N}\frac{h_k^j}{2}\left((\delta_i^j + \frac{g_k^j}{h_k^j})^2 - (\frac{g_k^j}{h_k^j})^2\right)$, we can organize Eq. 5 as below:

$$\sum_{j\in N} h_k^j(\delta_i^j + \frac{g_k^j}{h_k^j})^2 + 2\sum_{j\in M} g_k^j\delta_i^j - \sum_{j\in N}\frac{(g_k^j)^2}{h_k^j} - 2C = 0. \tag{6}$$

Note that, Eq. 6 above describes an elliptic paraboloid in p-dimensional space.

Analysis on the overlap between Ω and the derived privacy loss contour line. In the PGP scheme, we approximate the privacy loss based on Taylor expansion in the local region Ω around the current point ϕ_k, and derive the loss contour line to identify candidate points in the local region. Here we analyze the overlap between the local region Ω and the derived privacy loss contour line.

As analyzed in [30, 36], a single gradient descent step typically results in the update point (in our case denoted as $\phi'_k = \phi_k - \alpha g_k$) within the valid region Ω of Taylor approximation, i.e., the magnitude of the gradient update step αg_k is typically very small. Thus, the gradient descent point ϕ'_k usually stays well within the interior of Ω, and consequently, the loss contour line derived w.r.t. ϕ'_k naturally shares a significant overlap with the local region Ω. We also empirically observe that the approximation error (i.e., the error between the actual privacy loss reduction of the candidate points $L_p(\phi_k) - L_p(\phi_k + \delta_i)$ and the target loss reduction $L_p(\phi_k) - L_p(\phi'_k)$) is below 5%. This also implies that a non-negligible part of the contour line falls within the valid region Ω for Taylor approximation.

More details about the sampling. After deriving the contour line in Eq. 4 and Eq. 6, we sample candidate points over the contour line in the local region $\Omega = \{\delta_i \mid \|\delta_i\| \leq \varepsilon\}$, which approximately have the same loss value as ϕ'_k. To further ensure that the candidate points can lead to improved privacy protection ability, as mentioned in Sec. 3.1 in the main paper, we then check the actual privacy loss values of the candidate points and filter out the points that fail to achieve reduction in privacy loss.

D. More Details about the Training Pipeline and Algorithm

During the initialization, we compute utility gradients from the white-box surrogate model, and compute privacy gradients from the privacy evaluation model. We train the anonymization model by combining the utility gradients and privacy gradients (following [35]), to maintain the surrogate utility model's "satisfactory" performance while protecting privacy. The pre-trained white-box surrogate models are kept frozen in this process.

After initialization, we use GEP and PGP schemes to train the anonymization model with feedback (i.e., the final label output) from the black-box LMM and privacy gradients obtained from the privacy evaluation model. This algorithm for this training process is provided in Algorighm 1. Following the common pipeline of adversarial learning [6, 29, 35], when the privacy evaluation model's performance (i.e., accuracy of privacy attribute classification) drops below the threshold (0.95), we also update the privacy evaluation model toward stronger ability to classify the privacy attribute. During the training process, we randomly store updated parameters as archive points, and when the privacy loss reduction becomes minimal (less than $1e - 8$), we restart from one of the archive points following [26] to better explore the parameter space.

E. More Details about the Benchmarks and Metrics

More details about the benchmarks. We evaluate our framework following [6, 29, 35] on privacy-preserving action recognition (PPAR) benchmarks, HMDB51-VISPR [35] and UCF101-VISPR [35]. Specifically, we follow previous works [6, 35] to define the privacy attributes 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.

Besides PPAR benchmark, We also evaluate our framework with VQA task. To perform the "same-dataset" evaluation following PPAR benchmark on VQA task, we adopt the subcategories ("people and everyday life" and "sports and recreation") in OK-VQA dataset [23] with images that are most critical to privacy leakage and annotate the testing images with privacy attributes. Specifically, following [6, 35], the privacy attributes are 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 [14]. The annotated pri-

Algorithm 1 the Proposed PABP Framework

```
Require: black-box LMM $f_u$, the anonymization model $f_a$ with parameters $\phi$, initial radius $d$, learning rate $\alpha$.
$\phi \leftarrow$ GEP$(\phi, d)$.
for $K$ iterations do
    $\phi_0 \leftarrow \phi$
    while $J(\phi_k) = 0$ do
        $\alpha' \leftarrow \alpha$
        $\phi_{k+1}, \alpha' \leftarrow$ PGP$(\phi_k, \alpha')$
    end while
end for

function GEP$(\phi, d)$
    Initialized $\phi$.
    while $J(\phi) = 0$ do                                  ▷ $J(\phi)$ is defined in Eq. 1 in the main paper.
        for $m$ sampling points do
            Sample candidate point $\phi^i = \phi + \sigma^i$ and $\|\sigma^i\| = d$.
            if $J(\phi^i)$=1 then
                $\phi \leftarrow \phi^i$.
                return $\phi$
            end if
        end for
        $d \leftarrow 2d$
    end while
end function

function PGP$(\phi_k, \alpha)$
    Compute gradient $g_k = \nabla(L_p(\phi_k))$ given privacy loss function $L_p(\phi_k)$.
    Obtain privacy gradient descent point $\phi'_k = \phi_k - \alpha g_k$.
    if $J(\phi'_k) = 1$ then
        return $\phi'_k$
    else
        Approximate Hessian $H_k$ given $L_p(\phi_k)$ and obtain $m(\delta_i)$ (Eq. 2 in the main paper).
        Solve $m(\delta_i) = L_p(\phi'_k)$ (Eq. 4 in the main paper and Eq. 6 in this Supplementary.)
        for $m$ sampling points do
            Sample candidate point $\phi_k + \delta_i$ from the derived equation, where $\delta_i \in \Omega$ and $\Omega = \{\delta_i \mid \|\delta_i\| \leq \epsilon\}$.
            if $J(\phi_k + \delta_i) = 1$ then
                return $\phi_k + \delta_i, \alpha$
            end if
        end for
        return $\phi_k, \frac{\alpha}{2}$
    end if
end function
```

vacy attributes for each image are publicly available here[1] and the corresponding images can be downloaded from OK-VQA dataset[2].

Following [6, 29, 35], during training, in HMDB51-VISPR, the action recognition task is performed on the training set of HMDB51 dataset [12], while the privacy loss is obtained using the training set of VISPR dataset [28]. In UCF101-VISPR, during training, the utility task is performed using the training set of the UCF101 dataset [34], and the privacy protection loss is obtained using the VISPR training set. For the VQA experiments, following the similar pipeline in the PPAR benchmarks, during training, the VQA task is performed with the OK-VQA training data, and the privacy loss is obtained using the VISPR training data.

More details about the evaluation metrics. we follow [6, 13, 29, 35] to evaluate the performance of privacy

[1] https://github.com/phoebehxf/okvqa-privacy-attribute

[2] https://okvqa.allenai.org

protection using a privacy evaluation model (i.e., privacy attribute classifier) and adopt cMAP as the metric. More specifically, we follow [6] to adopt ResNet-50 [10] as the privacy evaluation model, and follow the same evaluation pipeline as [6] for the performance of privacy protection. The cMAP is calculated as follows [6]:

$$cMAP = \frac{1}{m} \sum_{i=1}^{m} \frac{TP_i}{TP_i + FP_i}, \tag{7}$$

where m is the number of privacy attribute classes, TP_i and FP_i are the numbers of true positives and false positives of the i-th privacy attribute class. The cMAP metric measures how well the privacy evaluation model can predict privacy information from the anonymized data, which reflects the performance of privacy preservation. A lower cMAP value indicates greater difficulty in recognizing privacy information, i.e., lower privacy leakage.

F. More Details about the Implementation.

More details about the models. Following previous works [6, 29, 35], we build the anonymization model based on U-Net [31]. To reduce the number of learnable parameters for tackling the problem involving the black-box model (as discussed in [27]), we make small modifications to the implementation of U-Net. Specifically, we implement the convolution layers following the lightweight depthwise separable convolution [4], and reduce the feature channels of the intermediate features following [21]. This reduces the number of parameters from 25M to $p = 9683$. For the privacy evaluation model (i.e., privacy attribute classifier), we follow [6, 13, 29] to adopt ResNet-50 [10].

More details about the implementation in training. In the training process, as introduced in Sec. 3.1 in the main paper, we initialize the anonymization model by adopting a small white-box surrogate utility model and training the anonymization model with gradients computed from the surrogate utility model and the privacy evaluation model. Here we provide the implementation details for this initialization. For action recognition experiments, we adopt the action recognition classifier R3D-18 [9] as the white-box surrogate utility model, which is pre-trained on UCF101 dataset [34] and HMDB51 dataset [12] following previous PPAR works [6, 29]. For VQA experiments, we adopt ConceptBert [8] as the white-box surrogate utility model, which is pre-trained on the OK-VQA dataset [23]. We set the initial radius in GEP as $d = 0.05$. In both GEP and PGP, the maximum number of sampled points from the sphere (or the loss contour line) is set to 15. We show the detailed training pipeline of the framework and the algorithm for the training process of the anonymization model in Algorithm 1.

More details about the evaluation. For evaluation of the LMMs on action recognition, we follow the retrieval-based evaluation [15] to obtain the action class from the LMM's output. We evaluate the LMM's performance on VQA following [3, 23].

G. Licenses

We use the VISPR dataset [28] following Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License. We use the HMDB51 dataset [12], UCF101 dataset [34], and OK-VQA dataset [23] following Creative Commons Attribution 4.0 International (CC BY 4.0) License.

We use GPT-4V [11] following the terms of using OpenAI services, and use Video-LLaVA [17] and LLaVA [19] following Apache License 2.0.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2

[2] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012. 1

[3] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 2, 7

[4] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 7

[5] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in neural information processing systems*, 27, 2014. 4

[6] 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, 6, 7

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[8] François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. ConceptBert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 489–498, Online, 2020. Association for Computational Linguistics. 7

[9] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Towards good practice for action recognition with spatiotemporal 3d convolutions. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2516–2521. IEEE, 2018. 1, 7

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[11] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 7

[12] 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. 6, 7

[13] Sudhakar Kumawat and Hajime Nagahara. Privacy-preserving action recognition via motion difference quantization. In *European Conference on Computer Vision*, pages 518–534. Springer, 2022. 6, 7

[14] 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*, pages 5106–5115, 2023. 5

[15] Shuailin Li, Yuang Zhang, Yucheng Zhao, Qiuyue Wang, Fan Jia, Yingfei Liu, and Tiancai Wang. Vlm-eval: A general evaluation on video large language models. *arXiv preprint arXiv:2311.11865*, 2023. 7

[16] TP Lillicrap. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015. 1

[17] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 7

[18] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989. 1

[19] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 7

[20] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720, 2023. 1, 4

[21] Zhanyu Ma, Dongliang Chang, Jiyang Xie, Yifeng Ding, Shaoguo Wen, Xiaoxu Li, Zhongwei Si, and Jun Guo. Fine-grained vehicle classification with channel max pooling modified cnns. *IEEE Transactions on Vehicular Technology*, 68(4):3224–3233, 2019. 7

[22] Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. *Advances in Neural Information Processing Systems*, 36:53038–53075, 2023. 1

[23] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 5, 7

[24] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417. PMLR, 2015. 1

[25] James Martens and Ilya Sutskever. Training deep and recurrent networks with hessian-free optimization. In *Neural Networks: Tricks of the Trade: Second Edition*, pages 479–535. Springer, 2012. 4

[26] Rafael Martí. Multi-start methods. *Handbook of metaheuristics*, pages 355–368, 2003. 5

[27] Changdae Oh, Hyeji Hwang, Hee-young Lee, YongTaek Lim, Geunyoung Jung, Jiyoung Jung, Hosik Choi, and Kyungwoo Song. Blackvip: Black-box visual prompting for robust transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24224–24235, 2023. 7

[28] 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, 6, 7

[29] Duo Peng, Li Xu, Qiuhong Ke, Ping Hu, and Jun Liu. Joint attribute and model generalization learning for privacy-preserving action recognition. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 5, 6, 7

[30] Haoxuan Qu, Lin Geng Foo, Yanchao Li, and Jun Liu. Towards more reliable confidence estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11): 13152–13169, 2023. 5

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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*, pages 234–241. Springer, 2015. 7

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[33] Sidak Pal Singh and Dan Alistarh. Woodfisher: Efficient second-order approximation for neural network compression. *Advances in Neural Information Processing Systems*, 33:18098–18109, 2020. 1, 4

[34] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6, 7

[35] 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, 2020. 2, 5, 6, 7

[36] Li Xu, Haoxuan Qu, Jason Kuen, Jiuxiang Gu, and Jun Liu. Meta spatio-temporal debiasing for video scene graph generation. In *European Conference on Computer Vision*, pages 374–390. Springer, 2022. 5