KOREA UNIVERSITY

Perturb a Model, Not an Image: Towards Robust Privacy Protection via Anti-Personalized Diffusion Models

Tae-Young Lee[1*] Juwon Seo[2*] Jong Hwan Ko[3†] Gyeong-Moon Park[1†]

[1]Korea University [2]Kyung Hee University [3]Sungkyunkwan University

Paper Code

NEURAL INFORMATION PROCESSING SYSTEMS

Presented by **Tae-Young Lee**

Visual & General Intelligence Lab.

01 Introduction

Motivation for our work

Text-to-Image Generation

- Recently, **Text-to-Image generation** has become a major research direction in generative AI.
  - Diffusion-based T2I models have rapidly evolved, achieving higher realism and controllability.

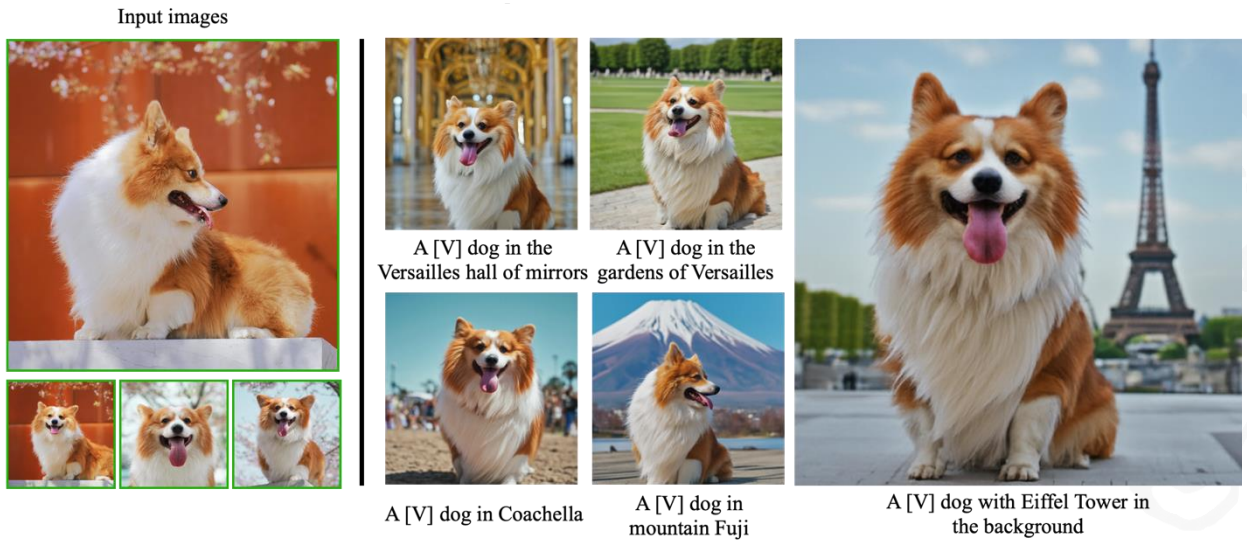
2022
Stable Diffusion 1.x

2023
Stable Diffusion 2.x/XL

2024
Stable Diffusion 3

Rise of Personalization

- In such a trend, **personalization methods** have emerged.



DreamBooth [1]

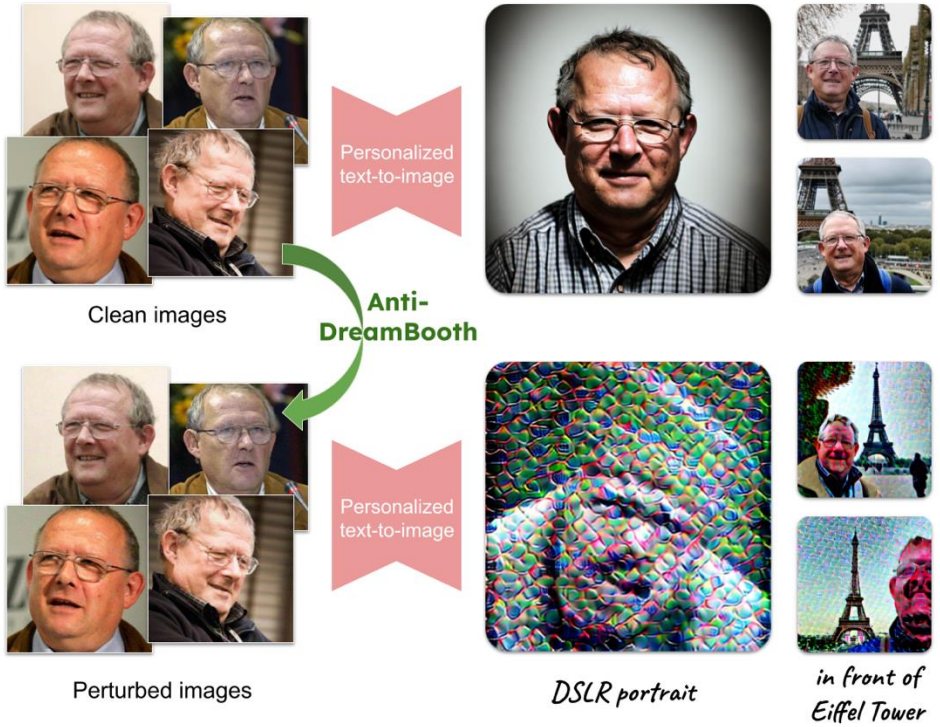
Custom Diffusion [2]

[1] Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." CVPR 2023.
[2] Kumari, Nupur, et al. "Multi-concept customization of text-to-image diffusion." CVPR 2023.

Privacy Concern of Personalization

- Despite their success, personalization methods also **raise some privacy concerns**.
  - Unauthorized content generation
  - Identity or likeness misuse (e.g., Deepfake)
  - Copyright infringement

- To counter these issues, researchers have aimed to **prevent unauthorized personalization**.
  - Based on the adversarial attack, they **add some perturbation** to the given images.
  - Attackers cannot personalize with these protected images.

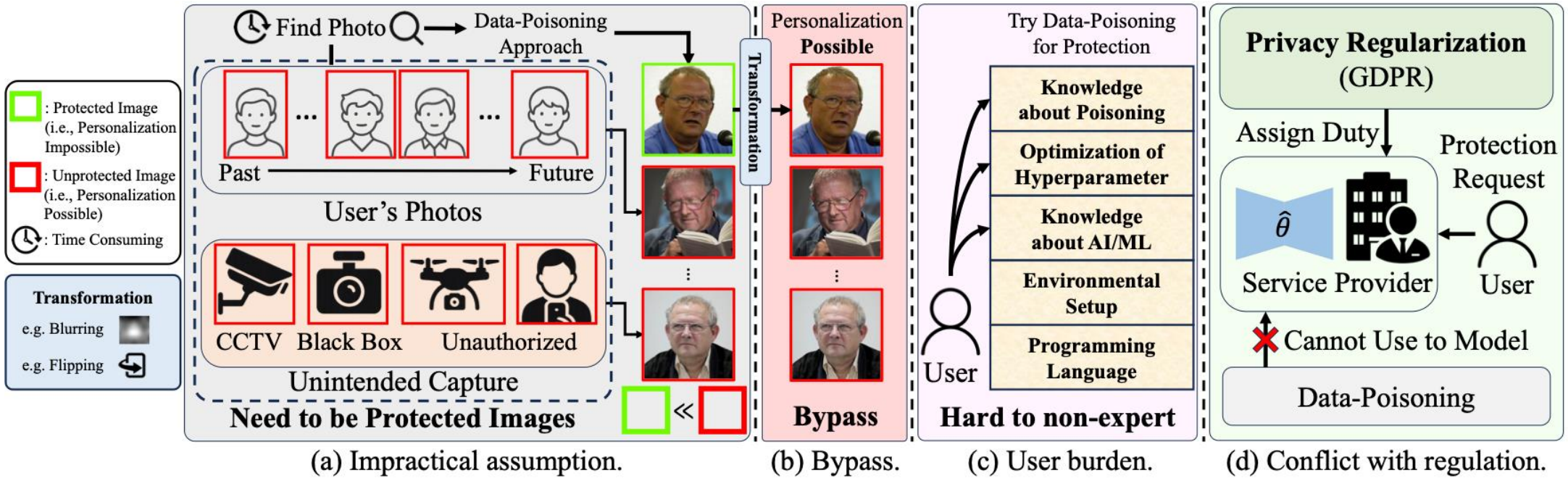


Anti-DreamBooth [3]

[3] Van Le, Thanh, et al. "Anti-dreambooth: Protecting users from personalized text-to-image synthesis." ICCV 2023.

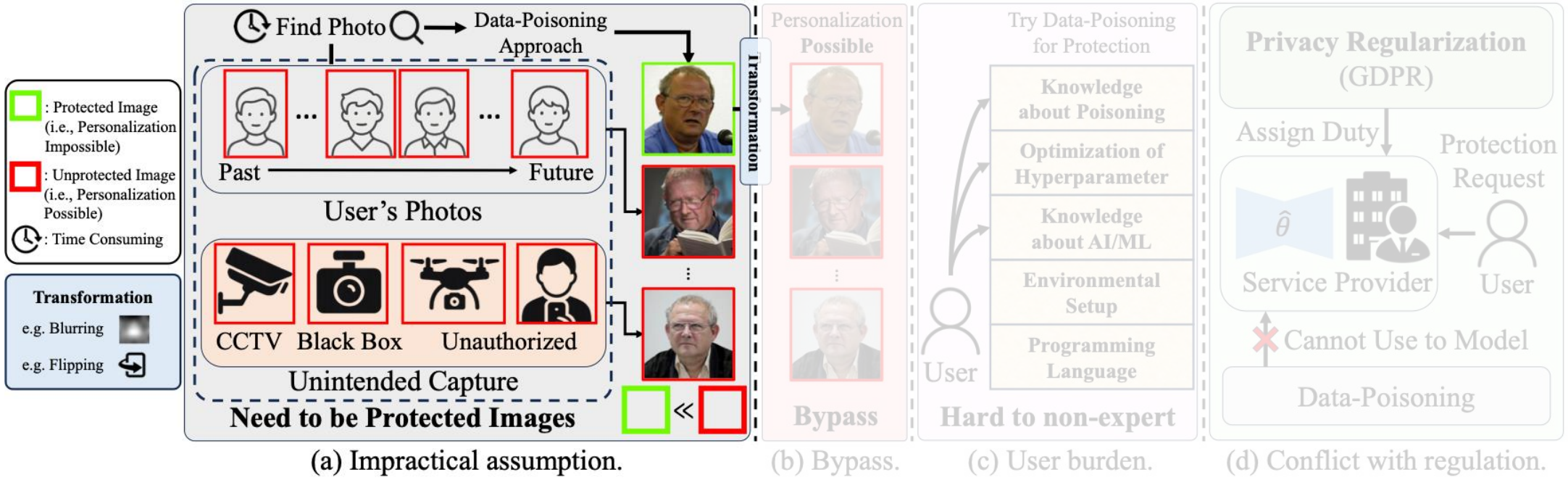
Motivation

- However, existing protection methods **only focused on data-level protection**.
  - Data-level protection modifies user data, but **fails to prevent personalization at the model level**.



Motivation

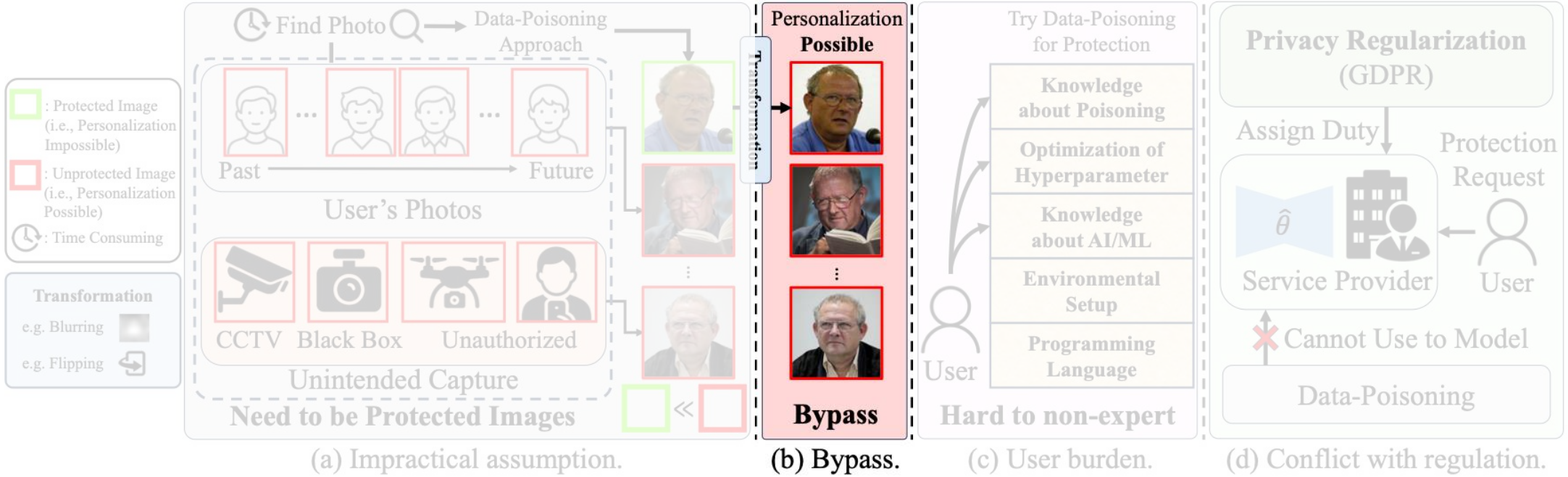
- **Impractical assumption**.
  - Cannot manage all images that contain the target subject.



(a) Impractical assumption. (b) Bypass. (c) User burden. (d) Conflict with regulation.

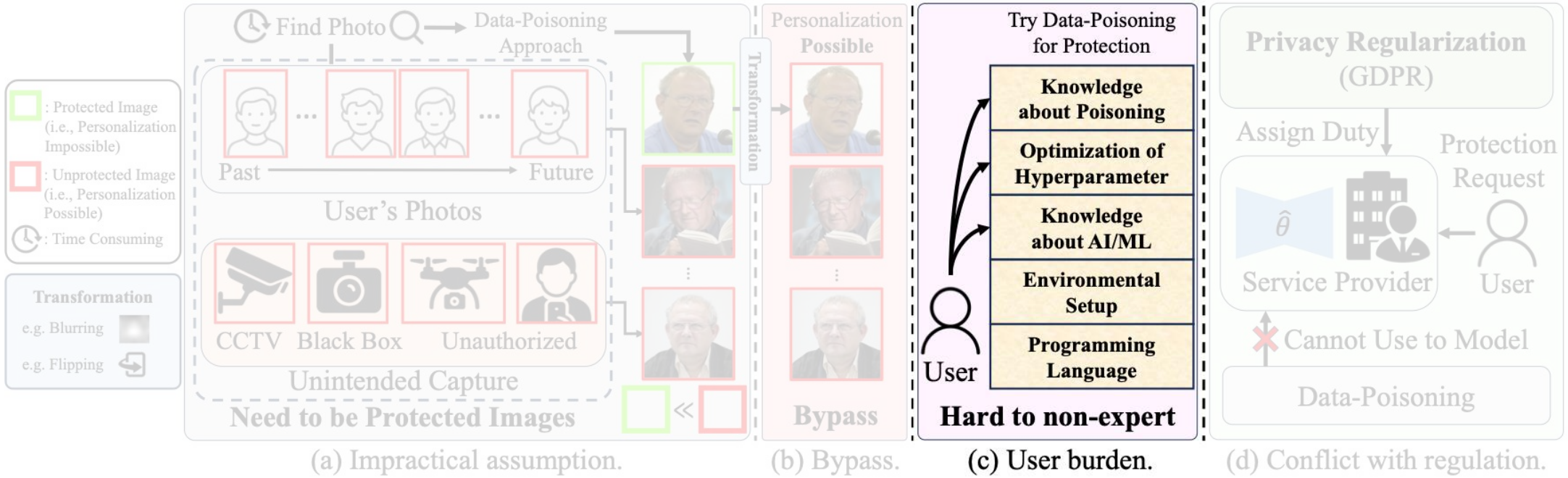
Motivation

- **Bypass**.
  - Easily bypassed by daily transformations (e.g., blurring, flipping).



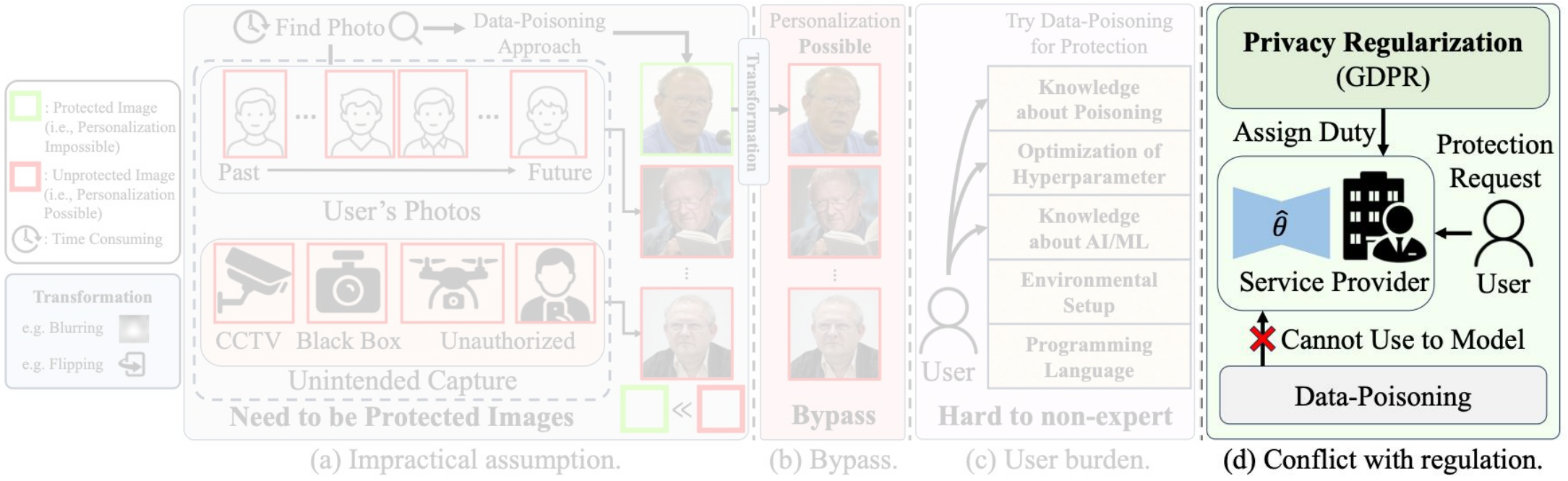
Motivation

- **User burden**.
  - Hard to apply for non-expert users.



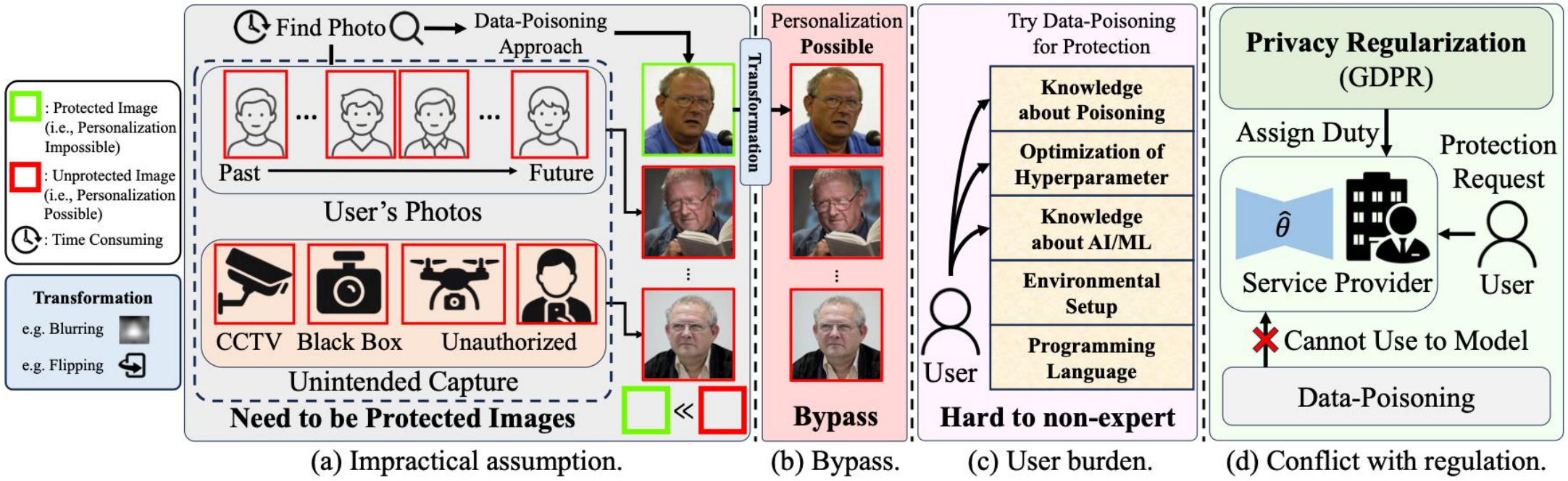
Motivation

- **Conflict with regulation**.
  - Service provider cannot use the existing approach on their service model.



Motivation

- **Key Insight**
  - Beyond the data dependency → **Perturb a Model, Not an Image**



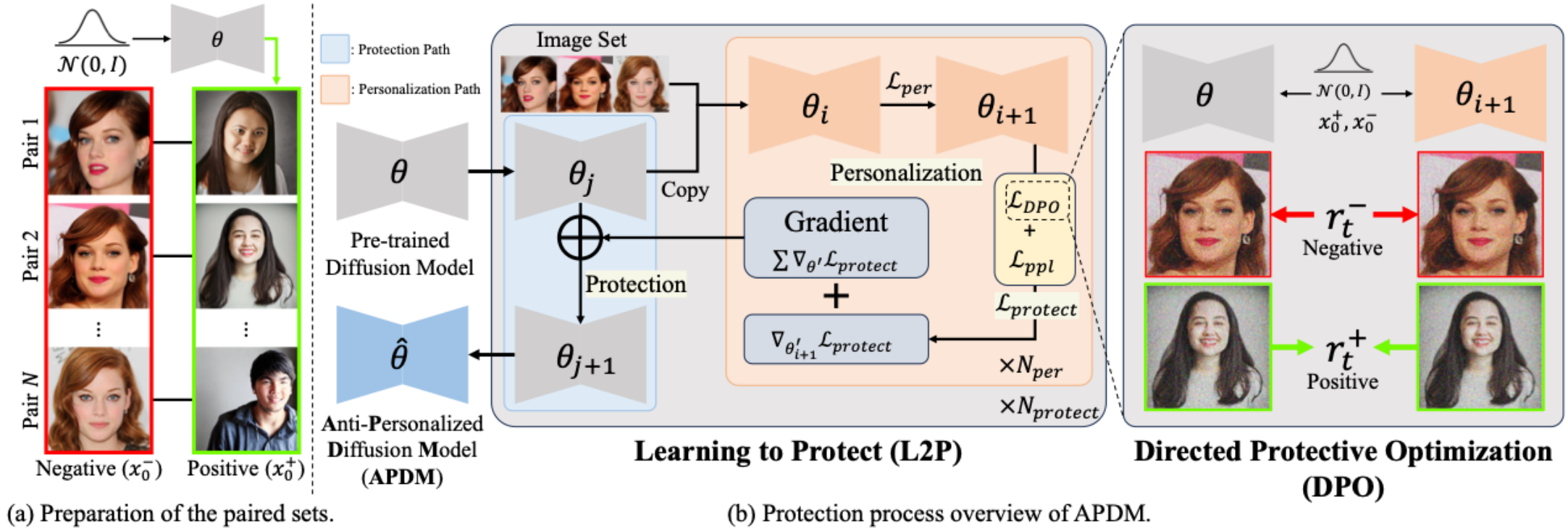
(a) Impractical assumption. (b) Bypass. (c) User burden. (d) Conflict with regulation.

02 APDM

Anti-Personalized Diffusion Model

Overview

- **Anti-Personalized Diffusion Model (APDM)**
  - **D**irected **P**rotective **O**ptimization (**DPO**)
  - **L**earning **to** **P**rotect (**L2P**)

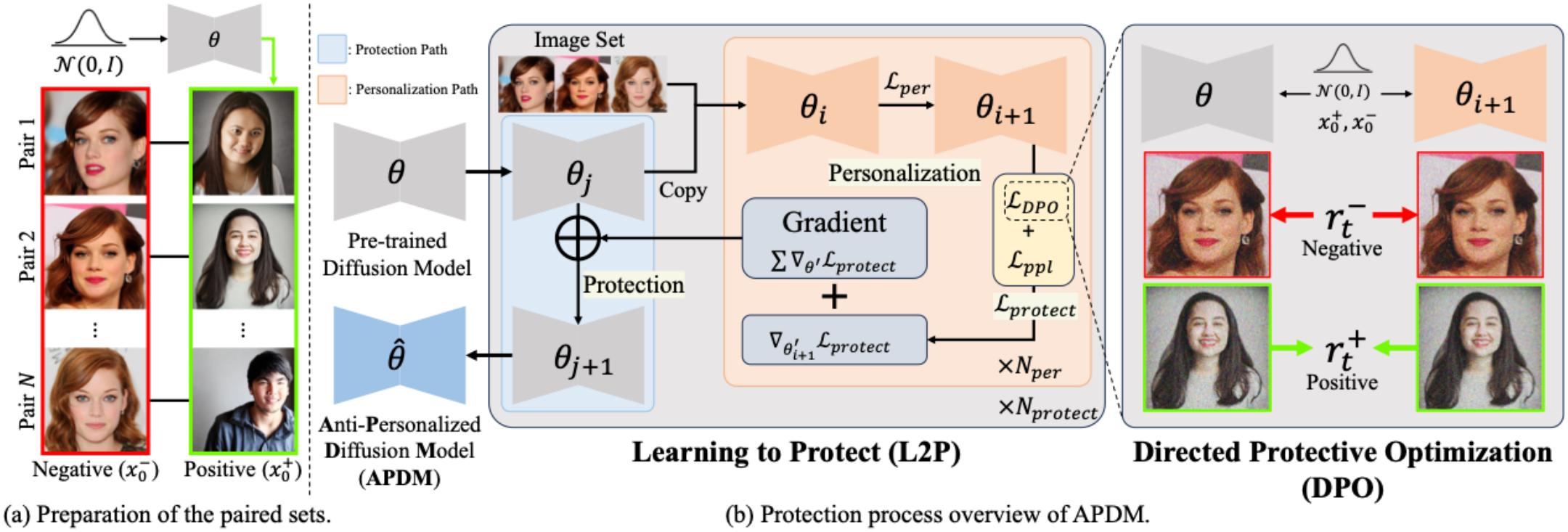


(a) Preparation of the paired sets.

(b) Protection process overview of APDM.

Directed Protective Optimization

- **D**irected **P**rotective **O**ptimization (**DPO**)
  - Inspired by Direct Preference Optimization [4], we directly guide the model on which information **should be learned** and **which should be suppressed**.



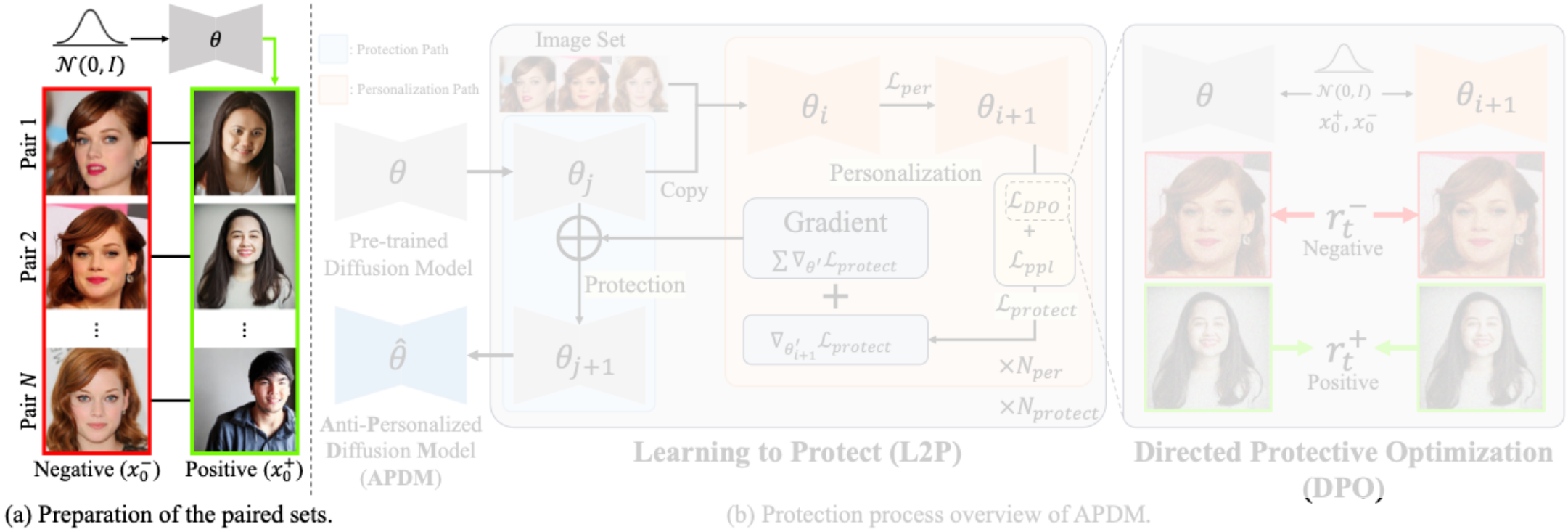
(a) Preparation of the paired sets.

(b) Protection process overview of APDM.

[4] Rafailov, Rafael, et al. "Direct preference optimization: Your language model is secretly a reward model." NeurIPS 2023.

Directed Protective Optimization

- Prepare the **paired sets** for DPO.
  - **Negative** (x_0^-): Images contain the **target of protection** (given).
  - **Positive** (x_0^+): Images contain the **encouraging results** after protection.
    - ✓ Generated by the T2I model.



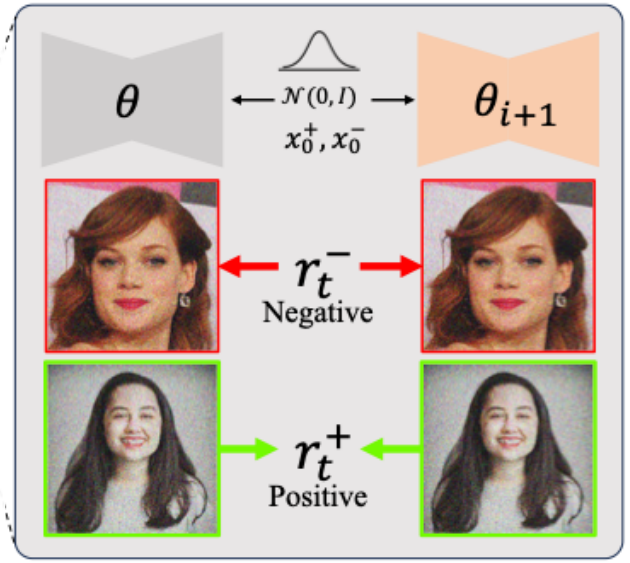
(a) Preparation of the paired sets.

(b) Protection process overview of APDM.

Directed Protective Optimization

- **DPO loss** function.
  - Goal: Encourages the generation of positive images while effectively suppressing the synthesis of negative images.

$$r^+ = \|\epsilon_\theta(x_t^+, t, c) - \epsilon\|_2^2 - \|\epsilon_\phi(x_t^+, t, c) - \epsilon\|_2^2,$$
$$r^- = \|\epsilon_\theta(x_t^-, t, c) - \epsilon\|_2^2 - \|\epsilon_\phi(x_t^-, t, c) - \epsilon\|_2^2,$$
$$\mathcal{L}_{DPO} = -\mathbb{E}_{x_0^+, x_0^-, c, t, \epsilon \sim N(0,1)} \log \sigma\big(-\beta(r^+ - r^-)\big).$$



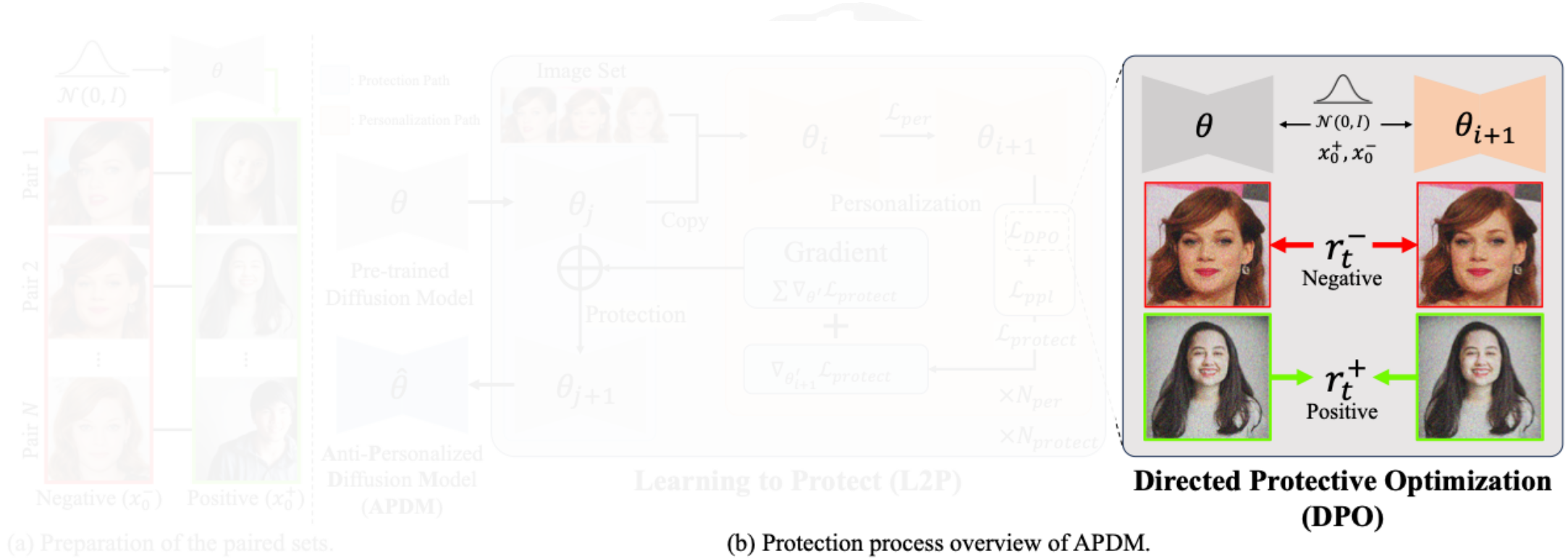
Directed Protective Optimization (DPO)

(b) Protection process overview of APDM.

Directed Protective Optimization

- Total protection loss function.
  - $\mathcal{L}_{ppl}$ is prior preservation loss for preserving the general knowledge about the target.

$$\mathcal{L}_{protection} = \mathcal{L}_{DPO} + \mathcal{L}_{ppl}.$$

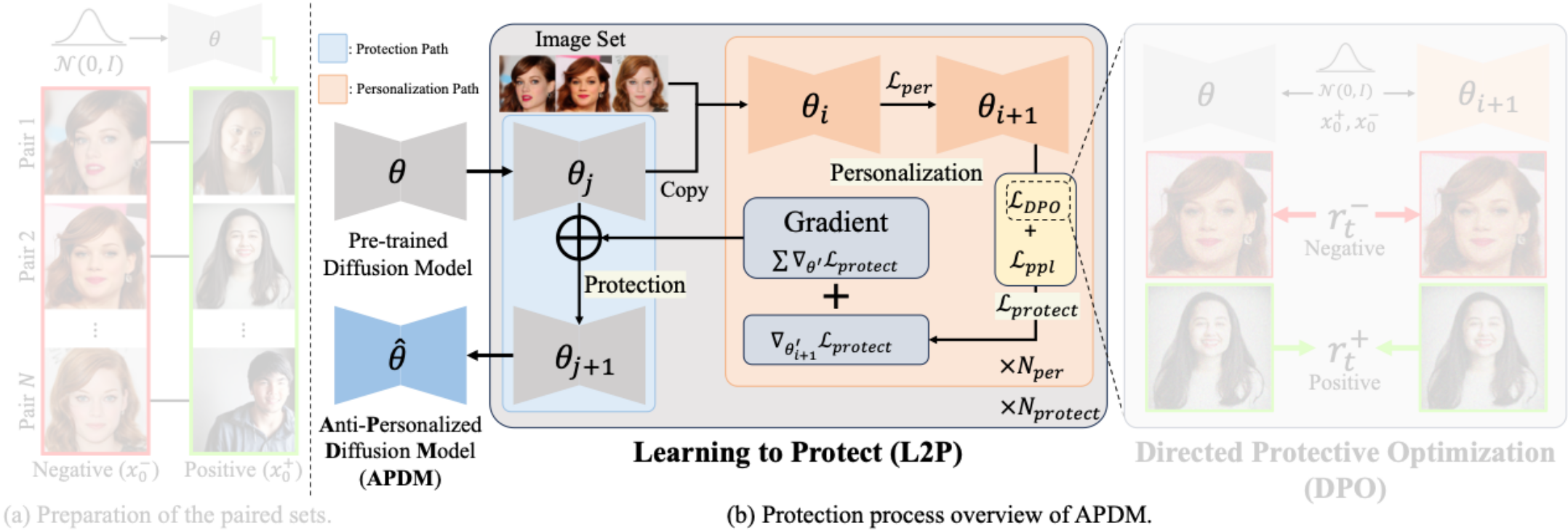


(a) Preparation of the paired sets.

(b) Protection process overview of APDM.

Learning to Protect

- **L**earning **to P**rotect (**L2P**).
  - Goal: **Maintain the protection** effect during personalization.
  - Approach: Accumulate **protection gradients** throughout the personalization path, and apply the aggregated gradient in the **protection path.**



(a) Preparation of the paired sets.

(b) Protection process overview of APDM.

03 Experiments

Setting & Results

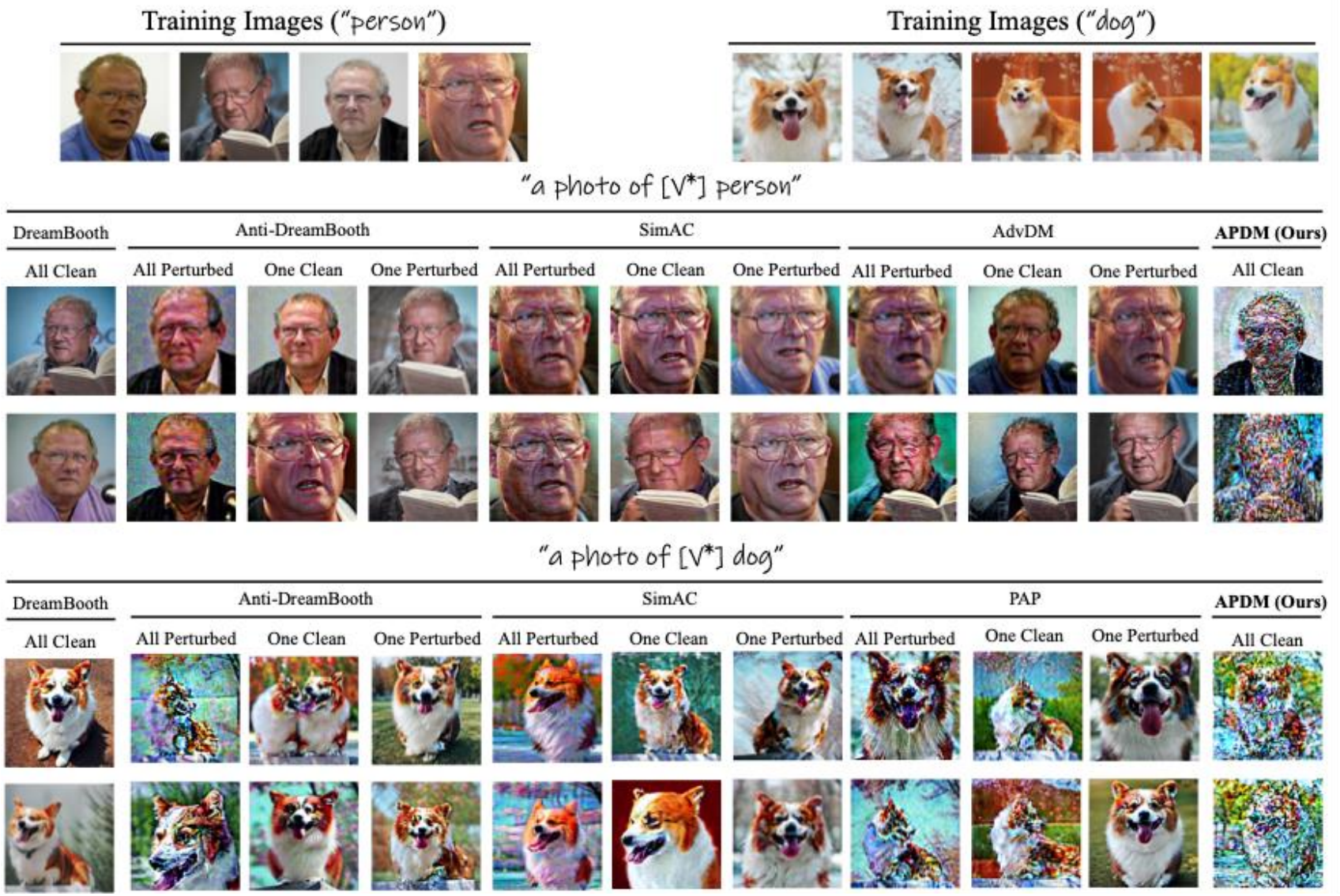
Experimental Setup

- **Metrics**
  - For protection performance:
    - ✓ DINO score (↓): Similarity-based metric.
    - ✓ BRISQUE (↑): Assessing image quality.
  - For the preservation of the pre-trained model's performance:
    - ✓ FID (↓): Overall image quality.
    - ✓ CLIP score (↑): Image-text alignment metric.
    - ✓ TIFA (↑): Image-text alignment metric.
    - ✓ GenEval (↑): Image-text alignment metric.
- **Datasets**
  - For person: CelebA-HQ and VGGFace2
  - For others: DreamBooth datasets.

Experimental Setup

- **Evaluation setting**
  - "# Clean Images" means the total number of clean (non-perturbed) images in the given set.
  - "# Clean Images": 0
    - ✓ Among the total N images, all images are perturbed images.
  - "# Clean Images": 1
    - ✓ Among the total N images, 1 is clean images and others are perturbed images.
  - "# Clean Images": $N-1$
    - ✓ Among the total N images, $N-1$ are clean images and other is perturbed image.
  - "# Clean Images": N
    - ✓ Among the total N images, N are clean images.

Experimental Results



Experimental Results

- Quantitative results.

Methods	# Clean Images	DINO (↓)			BRISQUE (↑)		
		"person"	*"dog"*	Avg.	*"person"*	*"dog"*	Avg.
DreamBooth [28]	N	0.6994	0.6056	0.6525	11.27	22.33	16.80
AdvDM [16]	0	0.5752	0.4247	0.4999	19.52	28.60	24.06
	1	0.5436	0.4393	0.4915	17.82	28.58	23.20
	$N-1$	0.6417	0.4775	0.5596	20.30	27.36	23.83
Anti-DreamBooth [30]	0	0.5254	0.4106	0.4680	26.90	30.23	28.56
	1	0.6081	0.4704	0.5393	23.76	27.49	25.63
	$N-1$	0.6951	0.5304	0.6127	15.48	25.26	20.37
SimAC [34]	0	0.4448	0.4374	0.4411	23.73	31.64	27.69
	1	0.5824	0.4537	0.5181	18.04	29.54	23.79
	$N-1$	0.6991	0.5370	0.6181	14.28	27.05	20.67
PAP [33]	0	0.6556	0.5120	0.5838	22.61	30.20	26.41
	1	0.6690	0.5032	0.5861	22.02	29.00	25.51
	$N-1$	0.7028	0.5270	0.6149	19.64	23.41	21.53
APDM (Ours)	N	**0.1375**	**0.0959**	**0.1167**	**40.25**	**60.74**	**50.50**

Experimental Results

- Quantitative results.

Methods	FID (↓)	CLIP (↑)	TIFA (↑)	GenEval (↑)
Stable Diffusion [27]	25.98	0.2878	78.76	0.4303
APDM (Ours)	28.85	0.2853	75.91	0.4017

04 Conclusion

Summary & Future Work

Conclusion

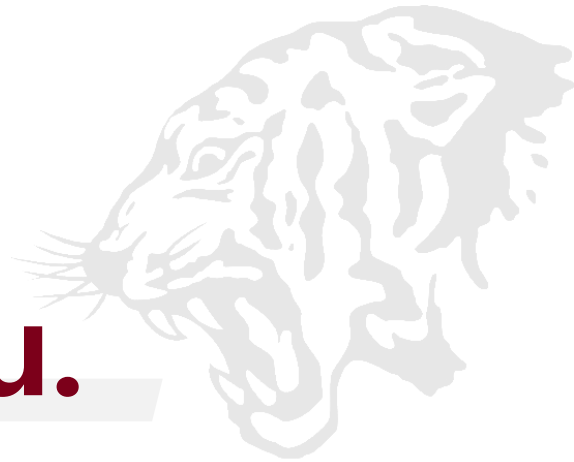
- Main task: **Robust Anti-Personalization**
  - Goal: Achieve protection that is *independent of given data* and can *counteract regulation*.
  - Approach: Move the protection target **from data to the model**.

- Propose framework: **APDM**
  - **D**irected **P**rotective **O**ptimization: Guides the model on what to suppress or preserve.
  - **L**earning **to** **P**rotect: Maintains the protection effect under continuous personalization.

- APDM achieves robust, data-independent protection with state-of-the-art performance.

- Future Direction
  - Multi-subject Protection, Continual Protection.

Thank you.



Paper

Code

Website

https://vgi.korea.ac.kr/

E-mail

tylee0415@korea.ac.kr
gm-park@korea.ac.kr