Anti-DreamBooth: Protecting users from personalized text-to-image synthesis — Supplementary Material —

Thanh Van Le*[1], Hao Phung*[1], Thuan Hoang Nguyen*[1], Quan Dao*[1], Ngoc N. Tran†[2], Anh Tran[1]
[1]VinAI Research [2]Vanderbilt University
v.{thanhlv19, haopt12, thuannh5, quandm7, anhtt152}@vinai.io, ngoc.n.tran@vanderbilt.edu

Abstract

In this supplementary PDF, we present additional experiments of our defense methods on two benchmark datasets, VGGFace2 and CelebA-HQ. We demonstrate the efficacy of our methods under different settings and provide more qualitative results for a better visual demonstration. In addition, we showcase the practical application of our methods by successfully disrupting a commercial AI service named Astria. We also include our code, some perturbed images generated by our method, and the output of Dreambooth models that were trained on these data in the supplementary package for better replication of our experiments and for future research.

1. Additional quantitative results

In the main paper, we comprehensively analyzed ASPL's performance on the VGGFace2 dataset. Here, we provide additional quantitative results on the CelebA-HQ dataset. We also report extra results with FSMG, the second-best defense algorithm, on the convenient settings.

1.1. Ablation studies

Text-to-image generator version. We investigate the effectiveness of our defense methods across different versions of SD models, including v1.4 and v1.5.

As reported in Tab. 1, ASPL significantly decreases the identity scores (ISM) in CelebA-HQ, confirming its defense's effectiveness. Its scores, however, are not as good as in VGGFaces2. We can explain it by the fact that CelebA-HQ images are more constrained in pose and quality, reducing the diversity of the image set for DreamBooth and making their combined perturbation effect less severe.

As for FSMG, there is a similar pattern in all metrics on both VGGFace2 and CelebA-HQ, as presented in Tab. 2. FSMG provides a slightly weaker defense compared with ASPL, confirming our observation in the main paper.

Noise budget. We further examine the impact of noise budget η on FSMG and ASPL using SD v2.1 in Tabs. 3 and 4. As expected, increasing the noise budget leads to better defense scores, either with FSMG or ASPL and either in VGGFace2 or CelebA-HQ. Again, ASPL outperforms FSMG on most evaluation scores.

1.2. Adverse settings

In the main paper, we verified that our best protection method, i.e., ASPL, remained effective in VGGFace2 when some components of the target DreamBooth training were unknown, resulting in a disparity between the perturbation learning and the DreamBooth finetuning. Here we repeat those defense experiments but on the CelebA-HQ dataset to further confirm ASPL's effectiveness.

Model mismatching. As can be seen in Tab. 5, the ASPL approach still works effectively on CelebA-HQ in the cross-model settings. Furthermore, the ensemble approach demonstrates a superior performance on all measurements, the same as the observation on VGGFace2.

Term mismatching. In realistic scenarios, the term representing the target in training DreamBooth might vary differently. To demonstrate this problem, we report ASPL's performance when the term "sks" is changed to "t@t". As can be seen in Tab. 5, our method still provides an extremely low ISM score, guaranteeing user protection regardless of the term mismatching.

Prompt mismatching. This is the challenging setting when the attacker uses a prompt different from the one used in perturbation learning to train his/her DreamBooth model. In Tab. 5, though there is a drop in some metrics compared with the convenience settings, either the ISM or BRISQUE score remains relatively good. This evidence further assures that our approaches are robust to the prompt mismatching problem.

* Equal contributions.
† Work done while at VinAI.

1.3. Uncontrolled settings

We examine APSL in the uncontrolled settings on CelebaA-HQ (Tab. 6) and observe the same trend as reported on the VGGFace2 dataset.

2. Real-world test.

In previous tests, we conducted experiments in laboratory mode. In this section, we examine if our proposed defense actually works in real-world scenarios by trying to disrupt personalized generation outputs of a black-box, commercialized AI service. We find Astria [1] satisfies our criteria and decide to use it in this test. Astria uses the basic DreamBooth setup that allows us to upload images of a specific target subject and input a generation prompt to acquire corresponding synthesized images. It also supports different model settings; we pick the recommended setting (SD v1.5 with face detection enabled) and a totally different one (Protogen 3.4 + Prism) for the tests.

We compare the output of Astria when using the original images and the adversarial images defended by our ASPL method with Stable Diffusion version 2.1 and $\eta = 0.05$ in Figs. 1 and 2, using two different subjects and with each model setting, respectively. As can be seen, our method significantly reduces the quality of the generated images in various complex prompts and on both target models. Even though these services often rely on proprietary algorithms and architectures that are not transparent to the public, our method remains effective against them. This highlights the robustness of our approach, which can defend against these services without requiring knowledge of their underlying configurations.

3. Qualitative results

We comprehensively analyzed our defense mechanism quantitatively in the main paper. Here, we provide additional qualitative results to back up those numbers and for visualization, as well.

3.1. Ablation studies

Text-to-image generator version. We compare the defense performance of ASPL using two different versions of SD models (v1.4 and v1.5) on VGGFace2 in Fig. 3. The output images produced by both models with both prompts are strongly distorted with notable artifacts. We observe the same behavior in the corresponding experiments on CelebA-HQ, visualized in Fig. 4.

Noise budget. In order to better understand the impact of the noise budget, we present a grid of images for ASPL on VGGFace2 where the upper bound of noise's magnitude increases along the vertical axis in Fig. 5. It is evident that when the noise budget increases, the visibility of noise becomes more pronounced. Moreover, the allocated noise budget heavily influences the degree of output distortion, resulting in a trade-off between the visibility of noise in perturbed images and the level of output distortion. For further visulization on CelebA-HQ, please refer to Fig. 6.

3.2. Adverse Setting

Model mismatching. In this section, we present the visual outputs of ASPL when a model mismatch occurs. Specifically, we train the image perturbation with SD v1.4, then use those images to disrupt DreamBooth models finetuned from v2.1 and v2.0, respectively. As illustrated in Fig. 7, our defense method is still effective in both cases, although transferring from v1.4 to v2.0 produces more noticeable artifacts than the previous scenario.

In addition to our primary analysis, our study provides qualitative results for E-ASPL, which employs an ensemble method to overcome the challenge of model mismatching. Specifically, we combined knowledge from three versions of SD models (v1.4, v1.5, and v2.1). The results, illustrated in Fig. 8, demonstrate the superior performance of E-ASPL in countering model mismatching where most images are heavily distorted.

Term mismatching. Despite the discrepancy of term replacement (from "sks" to "t@t"), ASPL still demonstrates its effectiveness on two provided subjects and two provided prompts (as in Fig. 9). However, the change in the training term may result in slightly weaker artifacts compared to the original setting.

Prompt mismatching. The results depicted in Fig. 9 indicate that the finetuning of the DreamBooth model with various prompts, such as "a DSLR portrait of *sks* person", can impact the degree of output distortion to some extent. It is important to note that prompt mismatching can alter the behavior of our defense method on a different prompt, such as "a photo of *sks* person", which can change the identity of the target subject in the generated images.

3.3. Uncontrolled settings.

All previous results are for controlled settings, in which we have access to all images needing protection. Here, we also include some qualitative results for uncontrolled settings where a mixture of clean and perturbed images are used for finetuning Dreambooth. We use the same settings as the one in the main paper, with the number of images for DreamBooth being fixed at 4 and the number of clean images gradually increase. As can be seen in Fig. 10, our method is more effective when more perturbed data are used and vice versa.

References

[1] Astria. https://www.astria.ai/. 2

Version	Defense?	"a photo of *sks* person"				"a dslr portrait of *sks* person"			
		FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑	FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑
v1.4	✗	0.07	0.48	0.66	16.09	**0.11**	0.40	0.67	10.31
	✓	**0.28**	**0.29**	**0.47**	**20.05**	0.06	**0.31**	**0.64**	**10.55**
v1.5	✗	0.06	0.53	0.69	14.45	**0.07**	0.39	0.68	8.95
	✓	**0.16**	**0.36**	**0.58**	**21.09**	0.06	**0.26**	**0.64**	**12.28**

Table 1: Defense performance of ASPL with different generator versions on CelebA-HQ in a convenient setting.

VGGFace2

Version	Defense?	"a photo of *sks* person"				"a dslr portrait of *sks* person"			
		FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑	FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑
v1.4	✗	0.05	0.46	0.65	21.06	0.08	0.44	0.64	10.05
	✓	**0.73**	**0.21**	**0.17**	**25.88**	**0.13**	**0.28**	**0.57**	**13.46**
v1.5	✗	0.07	0.49	0.65	18.53	0.07	0.45	0.64	10.57
	✓	**0.61**	**0.21**	**0.26**	**23.89**	**0.11**	**0.26**	**0.57**	**18.00**

CelebA-HQ

Version	Defense?	"a photo of *sks* person"				"a dslr portrait of *sks* person"			
		FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑	FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑
v1.4	✗	0.07	0.48	0.66	16.09	**0.11**	0.40	0.67	10.31
	✓	**0.29**	**0.32**	**0.48**	**20.83**	0.07	**0.29**	**0.63**	**12.00**
v1.5	✗	0.06	0.53	0.69	14.45	**0.07**	0.39	0.68	8.95
	✓	**0.13**	**0.38**	**0.60**	**20.43**	0.06	**0.28**	**0.65**	**13.27**

Table 2: Defense performance of FSMG with different generator versions on VGGFace2 and CelebA-HQ in a convenient setting.

VGGFace2

η	"a photo of *sks* person"				"a dslr portrait of *sks* person"			
	FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑	FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑
0	0.07	0.63	0.73	15.61	0.21	0.48	0.71	9.64
0.01	0.09	0.58	0.73	31.58	0.28	0.46	0.71	15.85
0.03	0.45	0.39	0.38	**37.82**	0.53	0.33	0.47	38.17
0.05^*	0.56	0.33	0.31	36.61	0.62	0.29	0.37	38.22
0.10	0.70	0.22	0.23	36.60	0.77	0.27	0.29	38.59
0.15	**0.77**	**0.20**	**0.20**	36.16	**0.83**	**0.22**	**0.26**	**39.17**

CelebA-HQ

η	"a photo of *sks* person"				"a dslr portrait of *sks* person"			
	FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑	FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑
0	0.10	0.68	0.72	17.06	0.26	0.44	0.72	7.30
0.01	0.12	0.68	0.73	19.55	0.30	0.46	0.71	6.60
0.03	0.15	0.57	0.71	33.89	0.27	0.41	0.73	22.67
0.05^*	0.34	0.48	0.56	36.13	0.35	0.36	0.66	33.60
0.10	0.73	0.32	0.27	**39.16**	0.67	0.24	0.43	**38.99**
0.15	**0.77**	**0.29**	**0.26**	38.22	**0.73**	**0.23**	**0.35**	38.22

Table 3: Defense performance of FSMG with different noise budgets on VGGFace2 and CelebA-HQ in a convenient setting. "*" is default.

η	"a photo of *sks* person"				"a dslr portrait of *sks* person"			
	FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑	FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑
0	0.10	0.68	0.72	17.06	0.26	0.44	0.72	7.30
0.01	0.11	0.67	0.72	19.97	0.27	0.45	0.72	6.65
0.03	0.12	0.60	0.71	34.34	0.25	0.44	0.73	18.29
0.05^*	0.31	0.50	0.55	38.57	0.34	0.39	0.63	34.89
0.10	0.73	0.36	0.30	**38.83**	0.74	0.27	0.36	**38.96**
0.15	**0.86**	**0.25**	**0.19**	38.67	**0.82**	**0.24**	**0.28**	38.86

Table 4: Defense performance of ASPL with different noise budgets on CelebA-HQ in a convenient setting. "*" is default.

	Train	Test	"a photo of *sks* person"				"a dslr portrait of *sks* person"			
			FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑	FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑
Model mismatch	v1.4	v2.1	0.37	0.48	0.53	39.28	0.34	0.39	0.64	33.50
	v1.4, 1.5, 2.1	v2.1	**0.39**	**0.46**	**0.48**	**38.25**	**0.44**	**0.34**	**0.57**	**37.29**
	v1.4	v2.0	0.40	0.46	0.51	38.88	0.43	0.36	0.60	22.21
	v1.4, 1.5, 2.1	v2.0	**0.56**	**0.43**	**0.43**	**41.83**	**0.55**	**0.33**	**0.51**	**29.93**
Term/ Prompt mismatch	DreamBooth prompt		"a photo of S_* person"				"a dslr portrait of S_* person"			
			FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑	FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑
	"*sks*" → "*t@t*"		0.20	0.17	0.64	26.49	0.17	0.10	0.65	1.14
	"a dslr portrait of *sks* person"		0.13	0.22	0.69	18.51	0.33	0.51	0.58	37.99

Table 5: Defense performance of ASPL on CelebA-HQ when the model, term, or prompt used to train the target DreamBooth model is different from the one used to generate defense noise. Here, S_* is "t@t" for the first row and "sks" for second row.

Perturbed	Clean	"a photo of *sks* person"				"a dslr portrait of *sks* person"			
		FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑	FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑
4	0	**0.31**	**0.50**	**0.55**	**38.57**	**0.34**	**0.39**	**0.63**	**34.89**
3	1	0.26	0.54	0.63	32.23	0.30	0.40	0.69	22.03
2	2	0.19	0.61	0.69	25.14	0.25	0.41	0.71	11.35
1	3	0.13	0.65	0.72	19.24	0.23	0.43	0.72	9.70
0	4	0.10	0.68	0.72	17.06	0.26	0.44	0.72	7.30

Table 6: Defense performance of ASPL on CelebA-HQ in uncontrolled settings. We include two extra results with 0 clean image (convenient setting) and 0 perturbed image (no defense) for comparison.

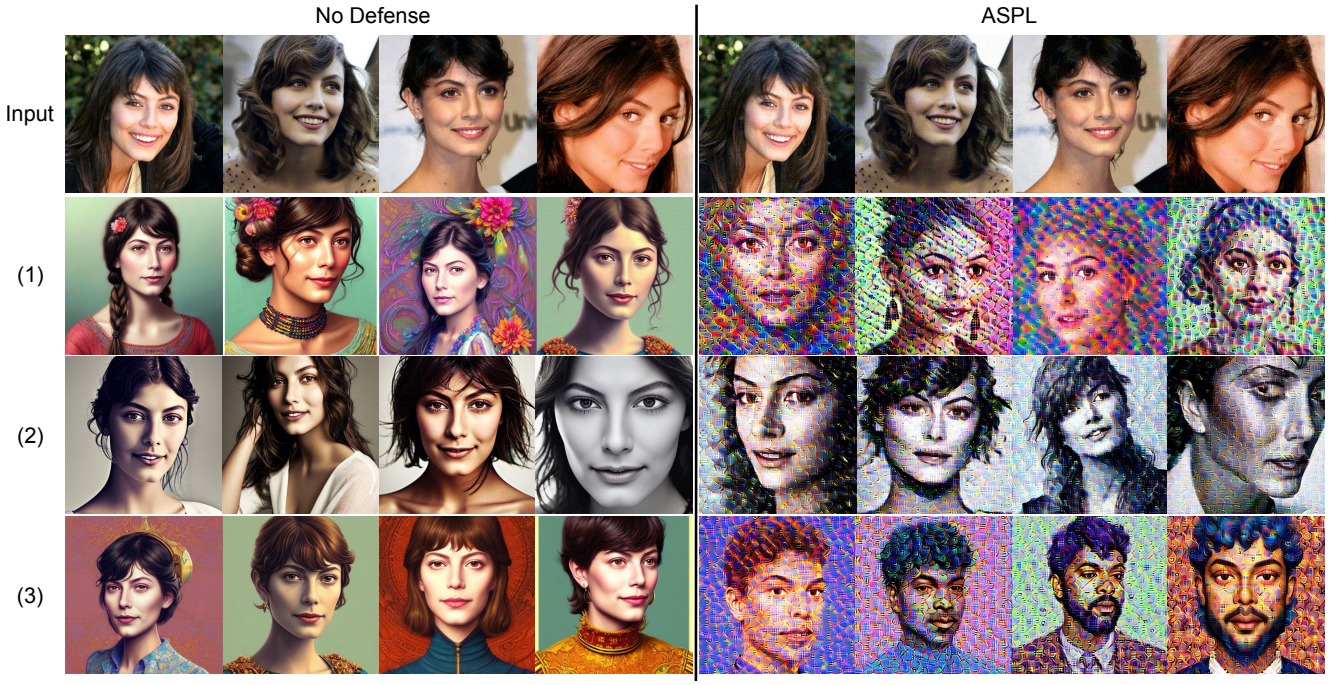

Figure 1: **Disrupting personalized images generated by Astria (SD v1.5 with face detection enabled)**. The prompts for image generation include: (1) "portrait of *sks* person portrait wearing fantastic Hand-dyed cotton clothes, embellished beaded feather decorative fringe knots, colorful pigtail, subtropical flowers and plants, symmetrical face, intricate, elegant, highly detailed, 8k, digital painting, trending on pinterest, harper's bazaar, concept art, sharp focus, illustration, by artgerm, Tom Bagshaw, Lawrence Alma-Tadema, greg rutkowski, alphonse Mucha", (2) "close up of face of *sks* person fashion model in white feather clothes, official balmain editorial, dramatic lighting highly detailed", and (3) "portrait of sks person prince :: by Martine Johanna and Simon Stålenhag and Chie Yoshii and Casey Weldon and wlop :: ornate, dynamic, particulate, rich colors, intricate, elegant, highly detailed, centered, artstation, smooth, sharp focus, octane render, 3d"

Figure 2: **Disrupting personalized images generated by Astria (Protogen with Prism and face detection enabled)**. The prompts for image generation include: (1) "portrait of *sks* person portrait wearing fantastic Hand-dyed cotton clothes, embellished beaded feather decorative fringe knots, colorful pigtail, subtropical flowers and plants, symmetrical face, intricate, elegant, highly detailed, 8k, digital painting, trending on pinterest, harper's bazaar, concept art, sharp focus, illustration, by artgerm, Tom Bagshaw, Lawrence Alma-Tadema, greg rutkowski, alphonse Mucha", (2) "close up of face of *sks* person fashion model in white feather clothes, official balmain editorial, dramatic lighting highly detailed", and (3) "portrait of sks person prince :: by Martine Johanna and Simon Stålenhag and Chie Yoshii and Casey Weldon and wlop :: ornate, dynamic, particulate, rich colors, intricate, elegant, highly detailed, centered, artstation, smooth, sharp focus, octane render, 3d"

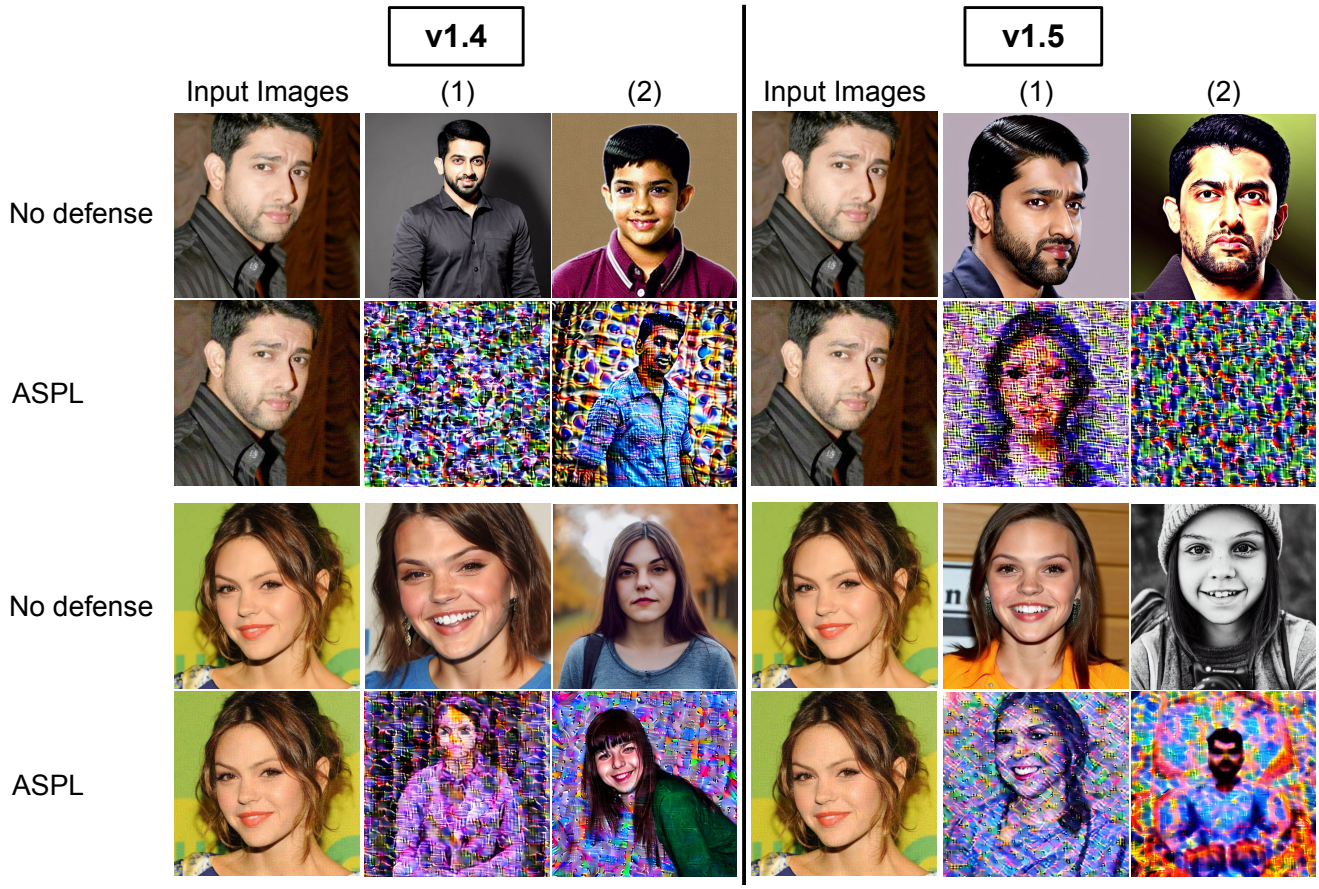

Figure 3: Qualitative results of ASPL with two different versions of SD models (v1.4 and v1.5) on VGGFace2. We provide in each test a single, representative input image. The generation prompts include (1) "a photo of *sks* person" and (2) "a dslr portrait of *sks* person".

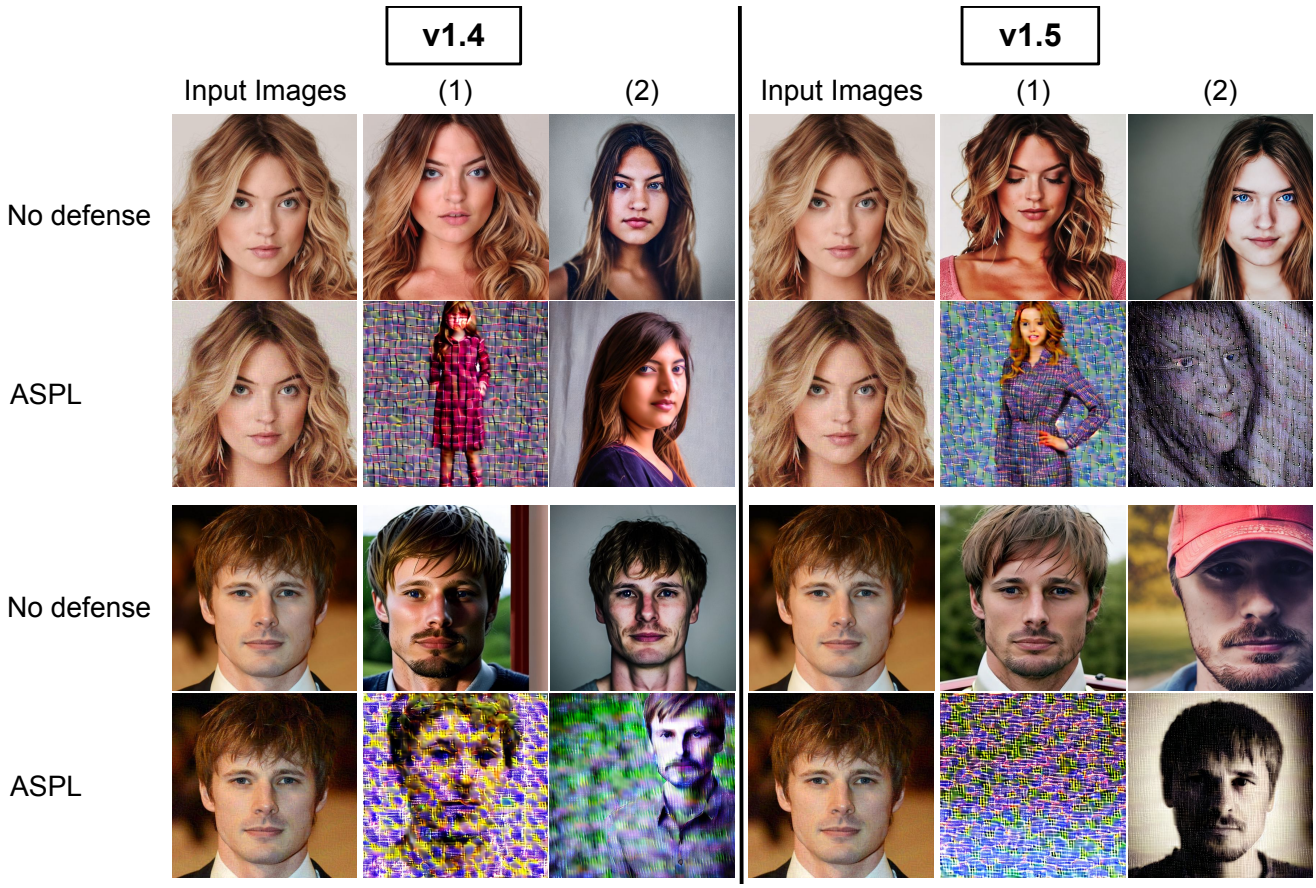

Figure 4: Qualitative results of ASPL with two different versions of SD models (v1.4 and v1.5) on CelebA-HQ. We provide in each test a single, representative input image. The generation prompts include (1) "a photo of *sks* person" and (2) "a dslr portrait of *sks* person".

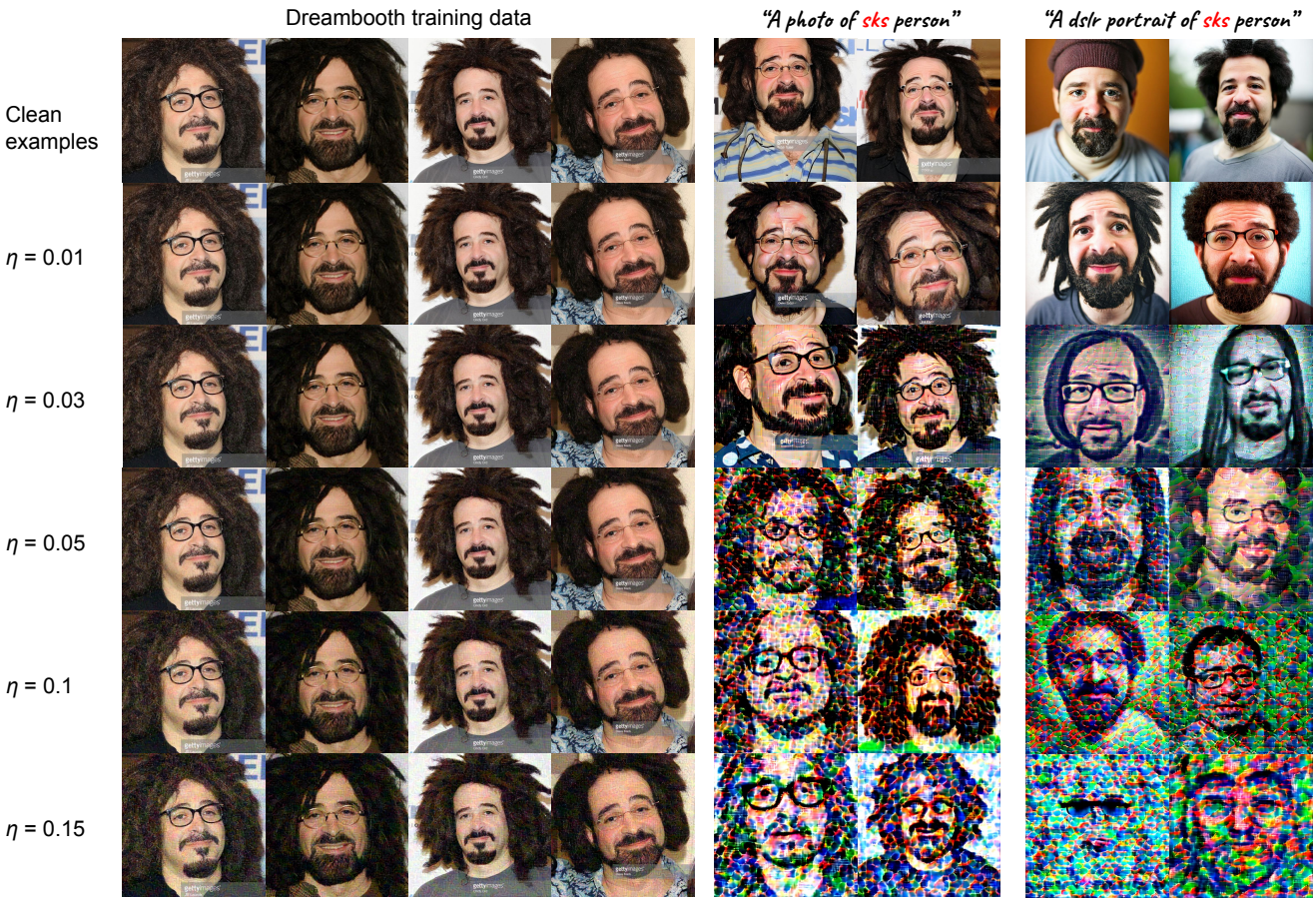

Figure 5: Qualativative results of ASPL with different noise budget on VGGFace2.

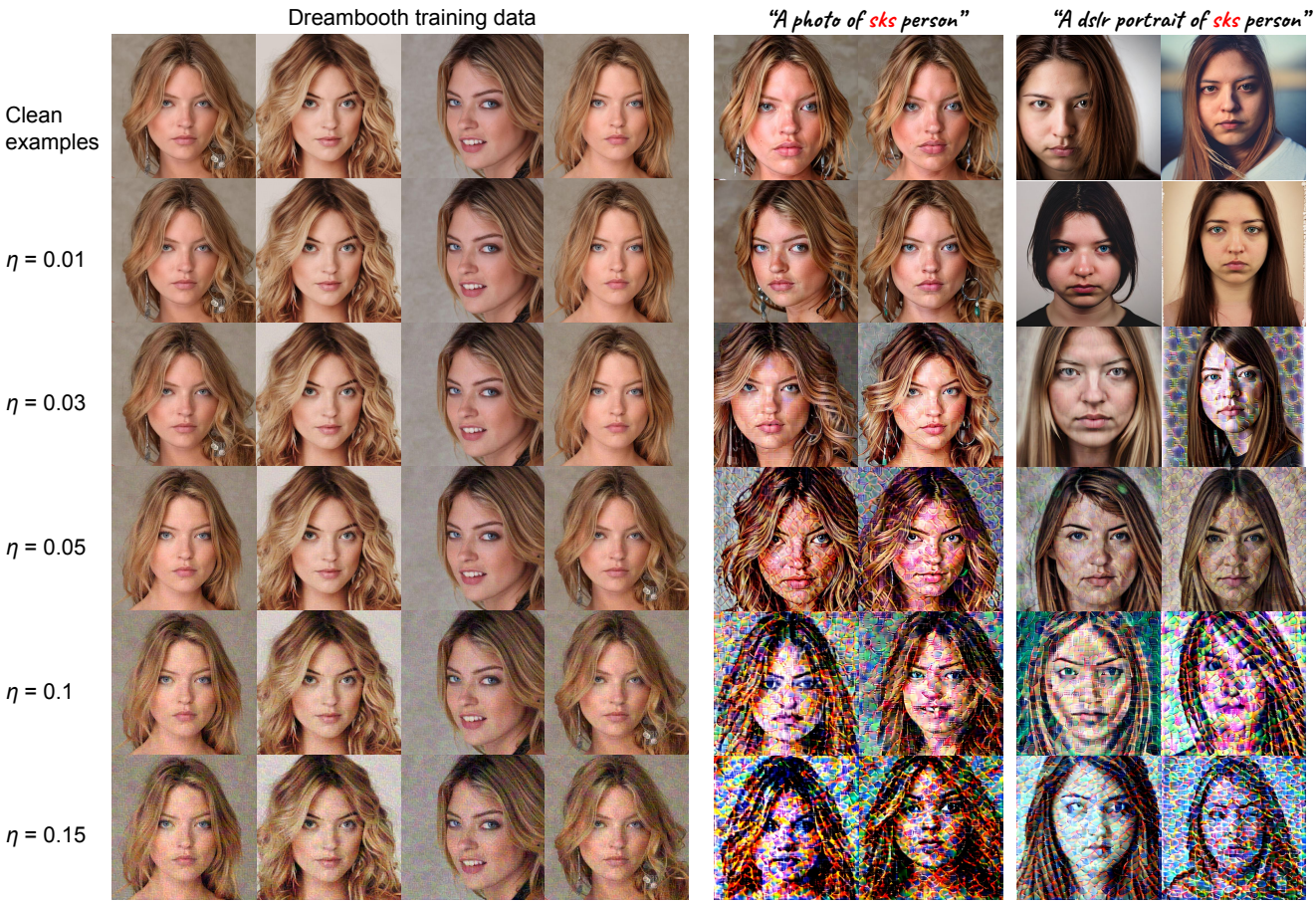

Figure 6: Qualitative results of ASPL with different noise budget on CelebA-HQ.

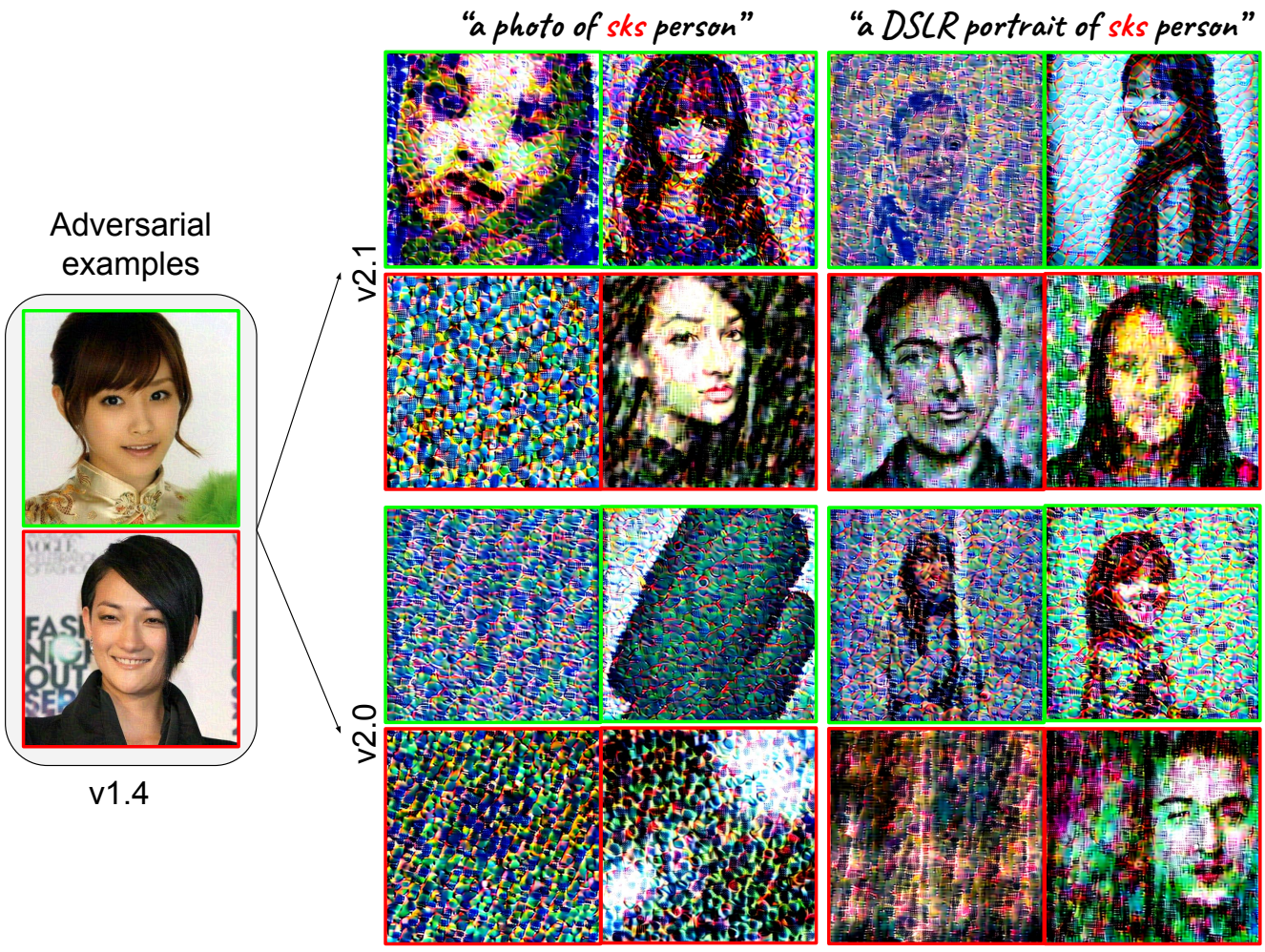

Figure 7: Qualitative results of ASPL in adverse settings on VGGFace2 where the SD model version in perturbation learning mismatches the one used in the DreamBooth finetuning stage (v1.4 $\rightarrow$ v2.1 and v1.4 $\rightarrow$ v2.0). We test with two random subjects and denote them in green and red, respectively.

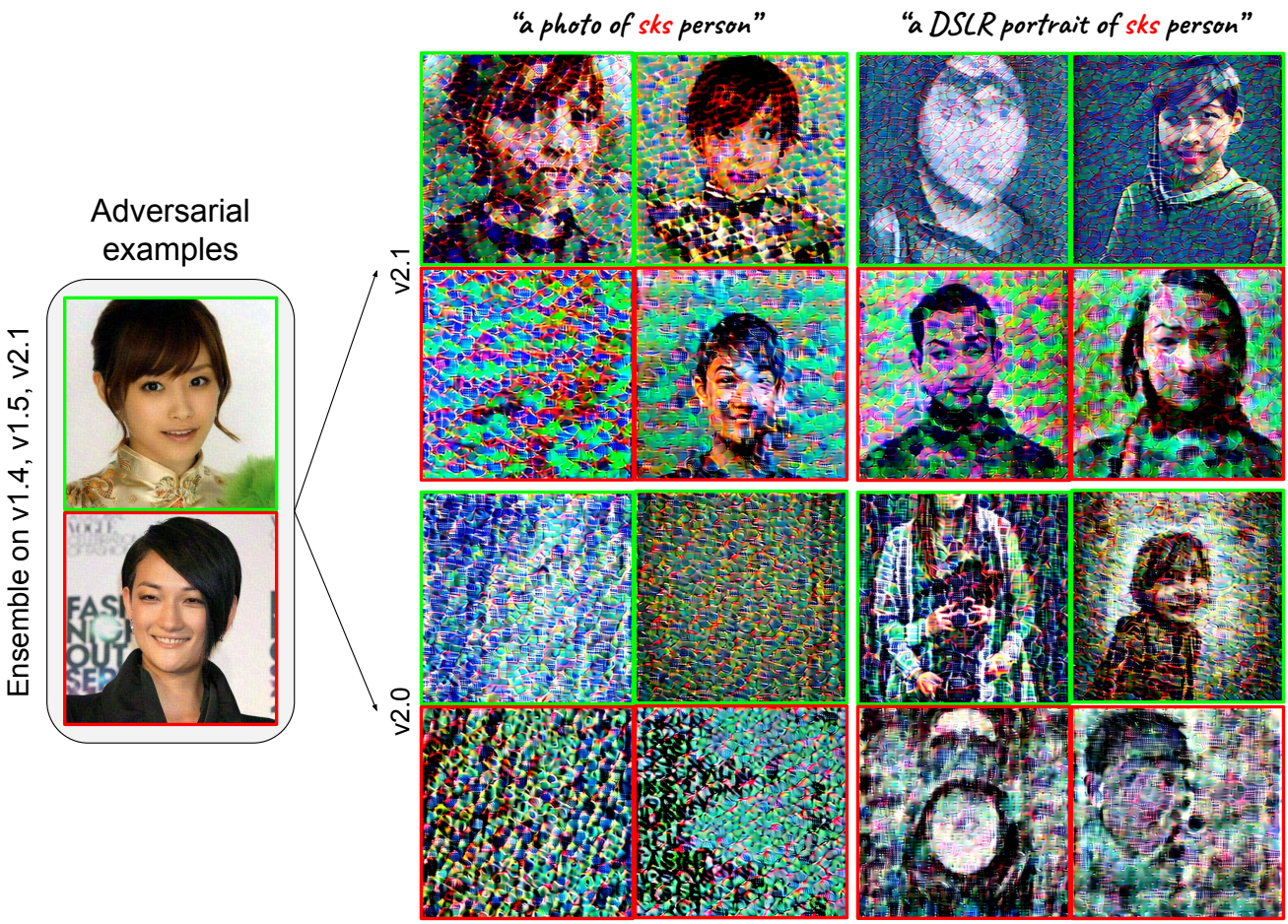

Figure 8: Qualitative results of E-ASPL on VGGFace2, where the ensemble model combines 3 versions of SD models, including v1.4, v1.5, and v2.1. Its performance is validated on two DreamBooth models finetuned on SD v2.1 and v2.0, respectively. We test with two random subjects and denote them in green and red, respectively.

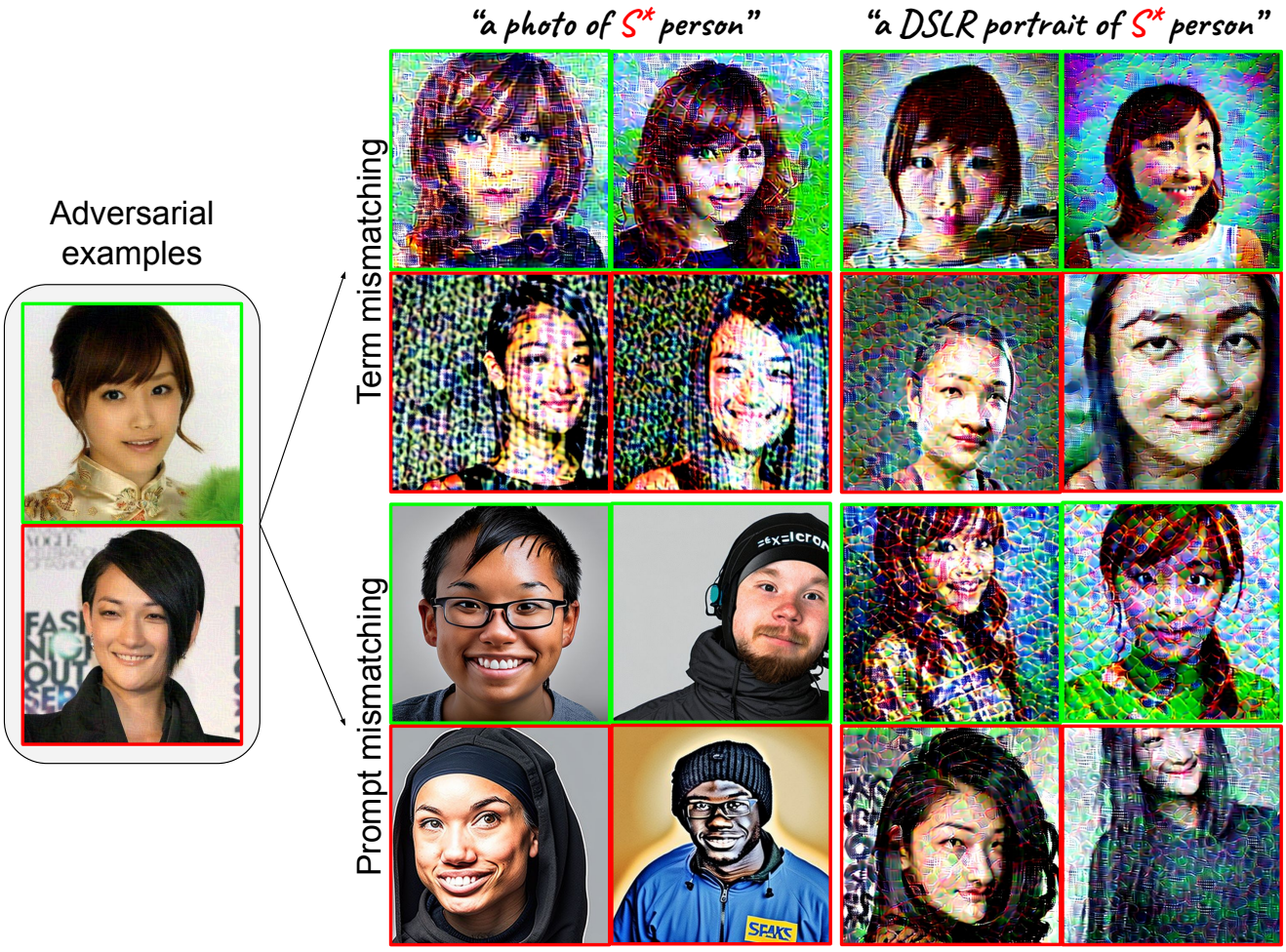

Figure 9: Qualitative results of ASPL on VGGFace2 where the training term and prompt of the target DreamBooth model mismatch the ones in perturbation learning. In the first scenario, the training term is changed from "sks" to "t@t". In the second scenario, the training prompt is replaced with "a DSLR portrait of *sks* person" instead of "a photo of *sks* person". Here, S^* is "t@t" for term mismatching and "sks" for prompt mismatching. We test with two random subjects and denote them in green and red, respectively.

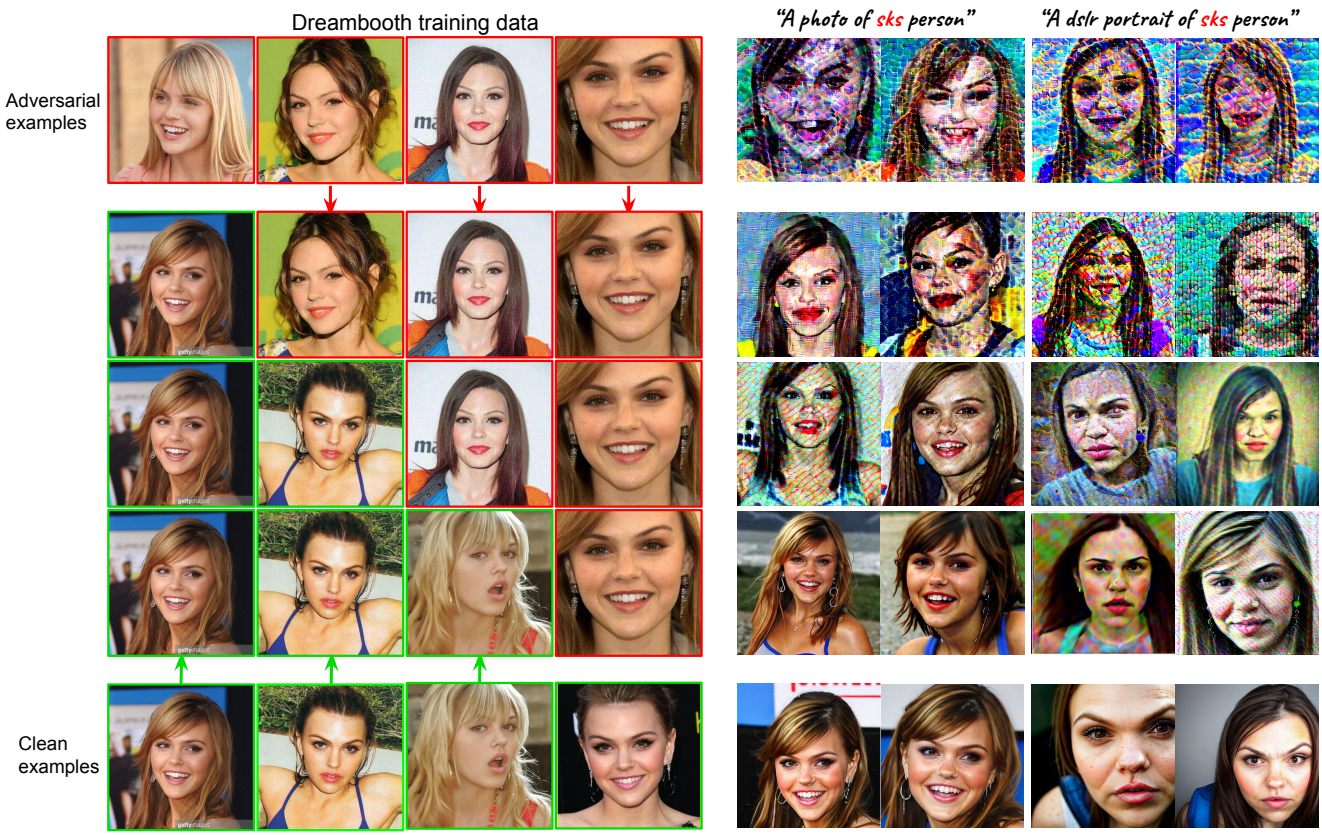

Figure 10: Qualitative results of ASPL in uncontrolled setting on VGGFace2. We denote the perturbed examples and the leaked clean examples in red and green, respectively.