JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

CLIP2Protect: Protecting Facial Privacy using Text-Guided Makeup via Adversarial Latent Search

Fahad Shamshad, Muzammal Naseer, Karthik Nandakumar

Session: THU-AM-392

MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

Motivation

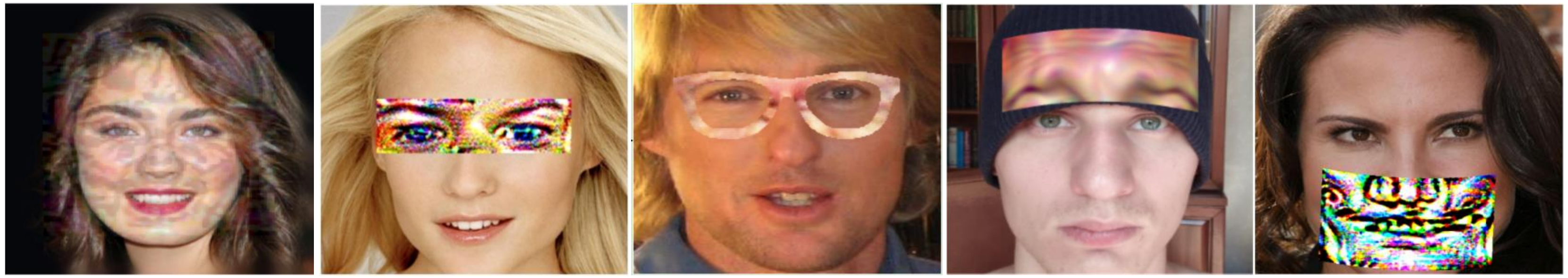
- Malicious **black-box** Face recognition (FR) systems pose a serious threat to personal security/privacy in **digital** world.

- Governmental and private entities can use FR systems **to track user activities** by scraping face images from social media platforms.

- With billions of people using social media, there is an urgent demand for effective privacy preservation methods.

Limitations of Existing Works

- Recent noise-constrained adversarial attacks to conceal user identity results in **artefacts**. Similarly, patch-based privacy approaches provide low privacy protection and **their large visible pattern** compromises naturalness.



- The closely related AMT-GAN needs a reference image for makeup style definition, which could affect practicality.
- Most methods focus on impersonating target identities, whereas the desired privacy objective is dodging.

Goals

- To propose a method that *i)* outputs ***naturalistic face images,*** ***ii)*** preserves the ***human perceived identity*** of user face, ***iii)*** ensures ***high privacy*** against ***black-box*** FR models, ***iv) offers user-friendliness***, and ***v)*** provides protection in both **impersonation and dodging** scenarios.

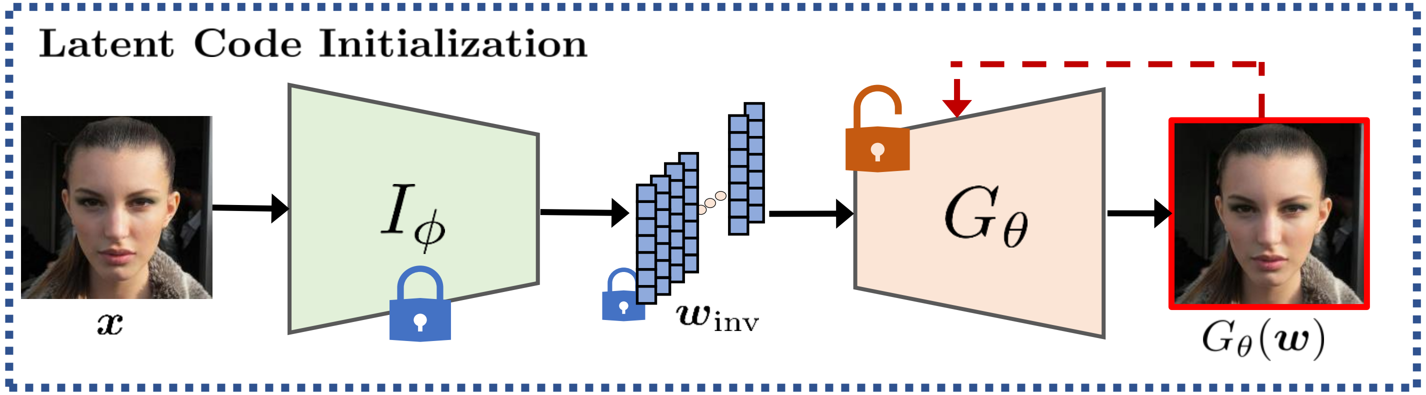
Contributions

- A two-step approach to find adversarial latent codes in a low-dimensional manifold of a pretrained generative model.

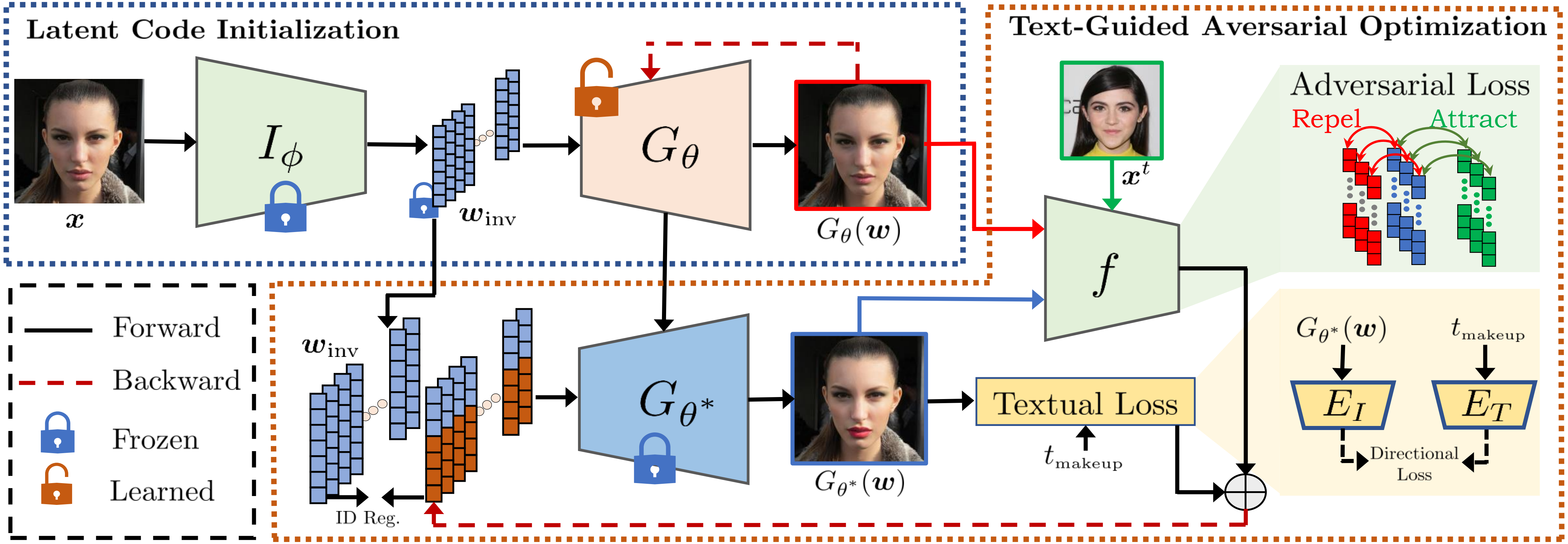
- User-defined textual (makeup) prompts to effectively hide attack information in the desired makeup style.

- A regularizer to preserve the identity-related attributes within the latent space of the generative model.

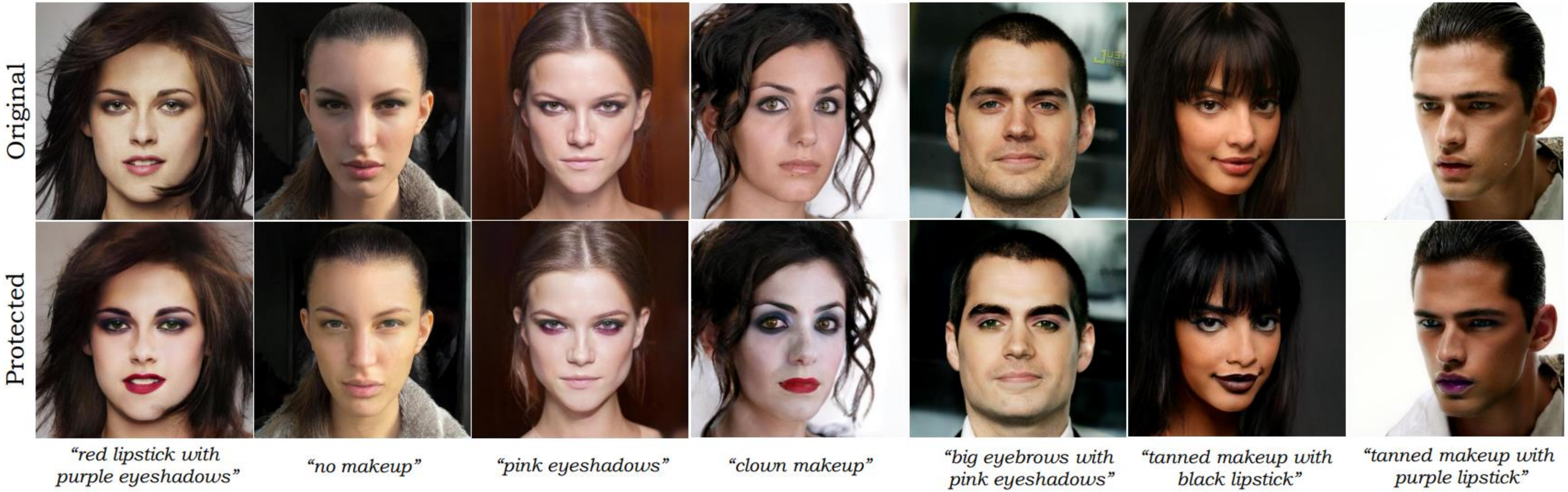
Overall Framework


Latent Code Initialization
$\boldsymbol{x}$
I_ϕ
$\boldsymbol{w}_{\text{inv}}$
G_θ
$G_\theta(\boldsymbol{w})$

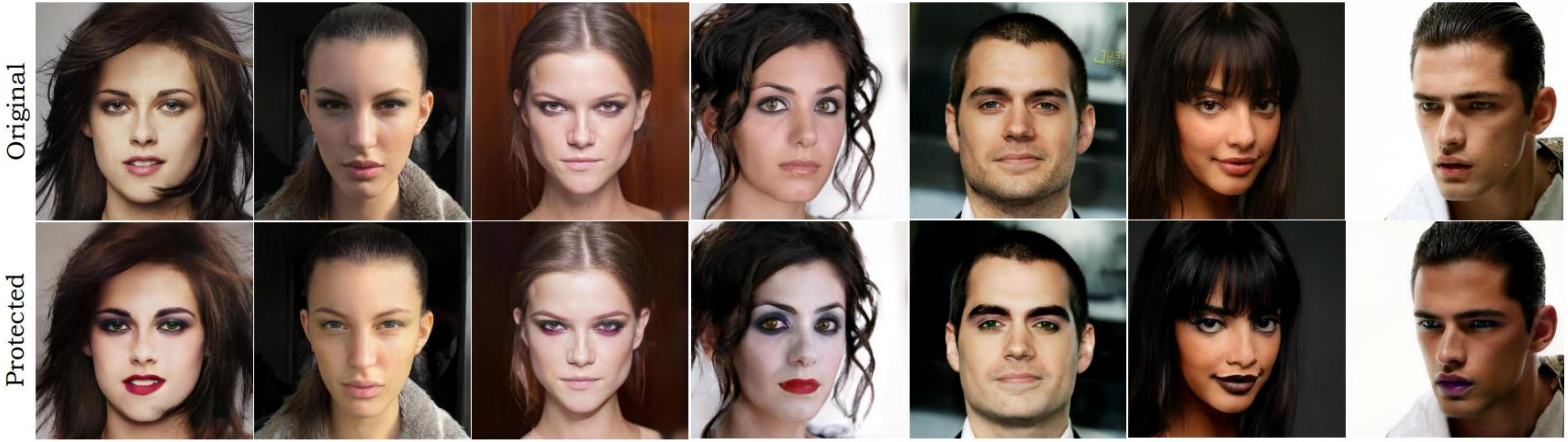
Overall Framework



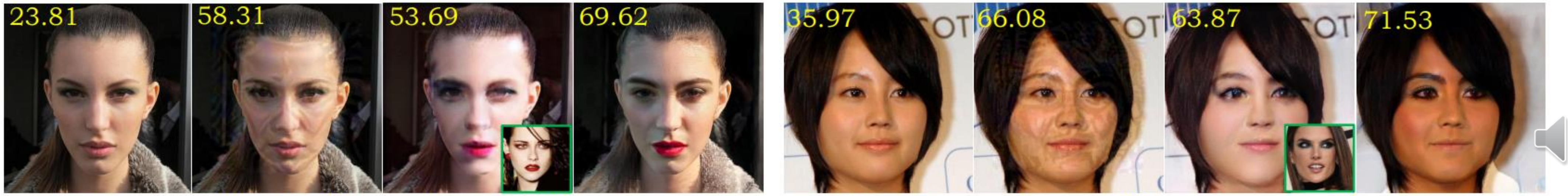
Qualitative Results


Original
Protected
"red lipstick with purple eyeshadows"
"no makeup"
"pink eyeshadows"
"clown makeup"
"big eyebrows with pink eyeshadows"
"tanned makeup with black lipstick"
"tanned makeup with purple lipstick"

Qualitative Results

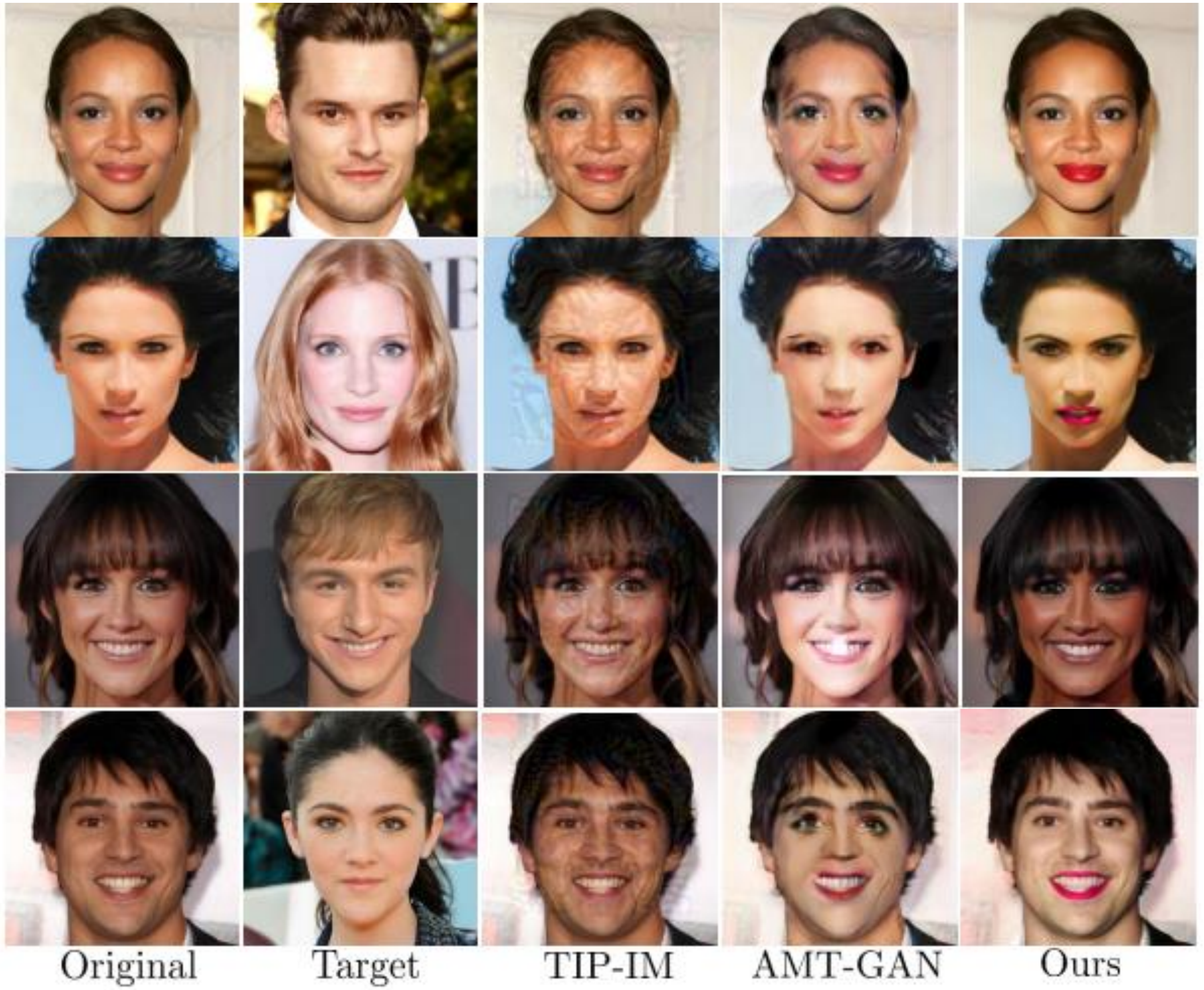


"red lipstick with purple eyeshadows" | "no makeup" | "pink eyeshadows" | "clown makeup" | "big eyebrows with pink eyeshadows" | "tanned makeup with black lipstick" | "tanned makeup with purple lipstick"



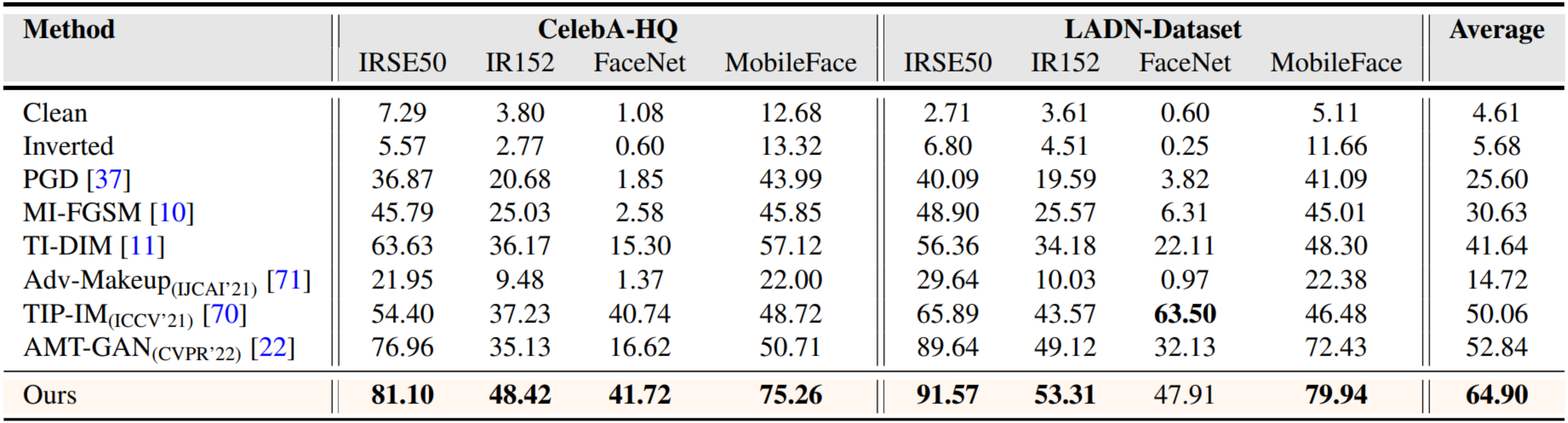
Original TIP-IM [70] AMT-GAN [22] Proposed Original TIP-IM [70] AMT-GAN [22] Proposed

Qualitative Results



Evaluation: Face Verification

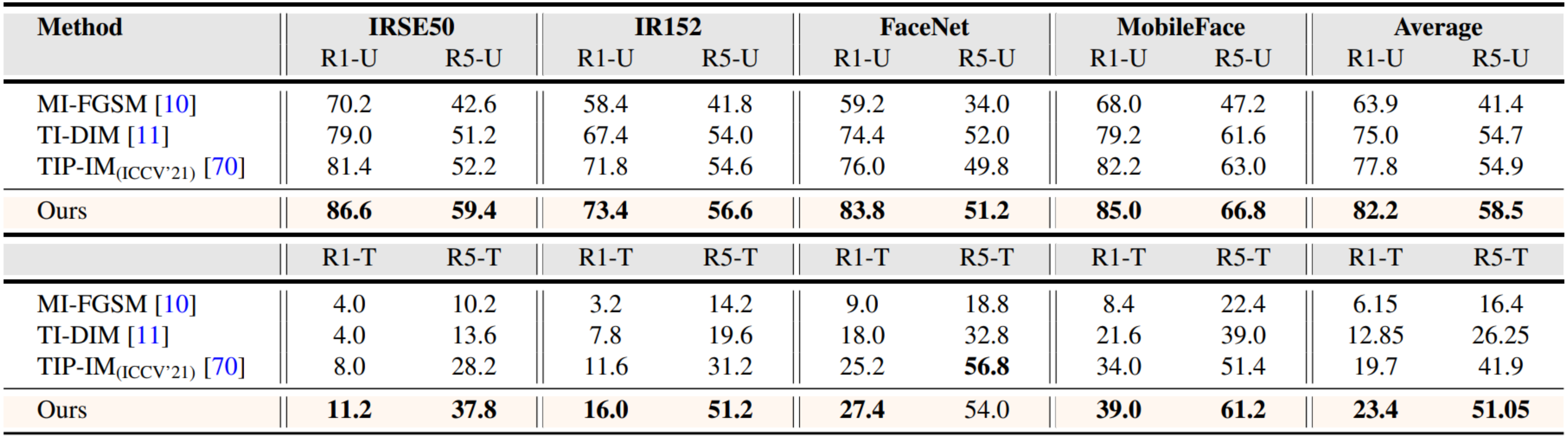
Protection success rate of black-box impersonation attack under the face verification task. For each column, the other three FR systems are used as surrogates to generate the protected faces.



Method	**CelebA-HQ**				**LADN-Dataset**				**Average**
	IRSE50	IR152	FaceNet	MobileFace	IRSE50	IR152	FaceNet	MobileFace	
Clean	7.29	3.80	1.08	12.68	2.71	3.61	0.60	5.11	4.61
Inverted	5.57	2.77	0.60	13.32	6.80	4.51	0.25	11.66	5.68
PGD [37]	36.87	20.68	1.85	43.99	40.09	19.59	3.82	41.09	25.60
MI-FGSM [10]	45.79	25.03	2.58	45.85	48.90	25.57	6.31	45.01	30.63
TI-DIM [11]	63.63	36.17	15.30	57.12	56.36	34.18	22.11	48.30	41.64
Adv-Makeup$_{\text{(IJCAI'21)}}$ [71]	21.95	9.48	1.37	22.00	29.64	10.03	0.97	22.38	14.72
TIP-IM$_{\text{(ICCV'21)}}$ [70]	54.40	37.23	40.74	48.72	65.89	43.57	**63.50**	46.48	50.06
AMT-GAN$_{\text{(CVPR'22)}}$ [22]	76.96	35.13	16.62	50.71	89.64	49.12	32.13	72.43	52.84
Ours	**81.10**	**48.42**	**41.72**	**75.26**	**91.57**	**53.31**	47.91	**79.94**	**64.90**

Evaluation: Face Identification

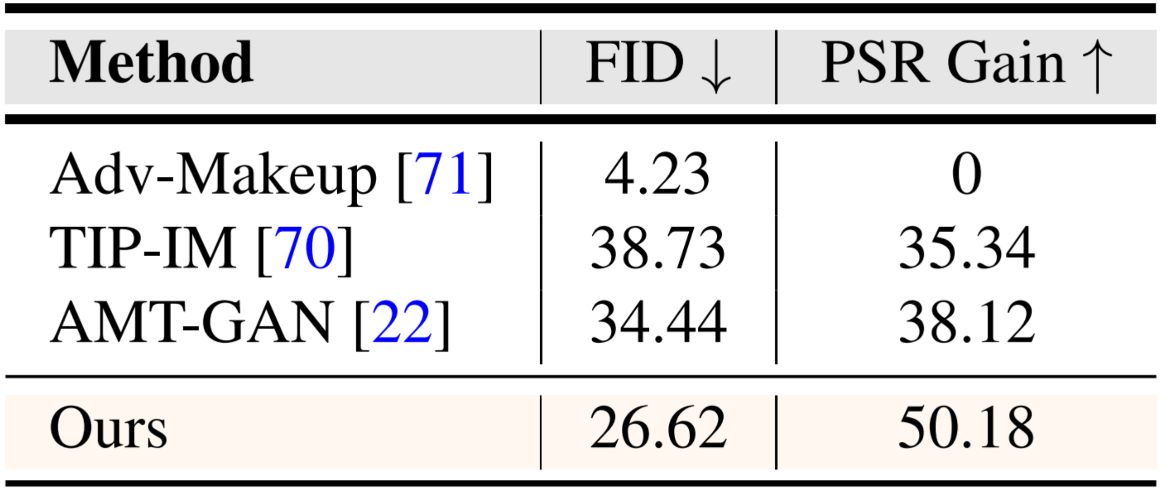
Protection success rate of black-box dodging (top) and impersonation (bottom) attacks under the face identification task. For each column, the other three FR systems are used as surrogates to generate the protected faces.



Method	IRSE50		IR152		FaceNet		MobileFace		Average	
	R1-U	R5-U	R1-U	R5-U	R1-U	R5-U	R1-U	R5-U	R1-U	R5-U
MI-FGSM [10]	70.2	42.6	58.4	41.8	59.2	34.0	68.0	47.2	63.9	41.4
TI-DIM [11]	79.0	51.2	67.4	54.0	74.4	52.0	79.2	61.6	75.0	54.7
TIP-IM$_{(ICCV'21)}$ [70]	81.4	52.2	71.8	54.6	76.0	49.8	82.2	63.0	77.8	54.9
Ours	**86.6**	**59.4**	**73.4**	**56.6**	**83.8**	**51.2**	**85.0**	**66.8**	**82.2**	**58.5**
	R1-T	R5-T	R1-T	R5-T	R1-T	R5-T	R1-T	R5-T	R1-T	R5-T
MI-FGSM [10]	4.0	10.2	3.2	14.2	9.0	18.8	8.4	22.4	6.15	16.4
TI-DIM [11]	4.0	13.6	7.8	19.6	18.0	32.8	21.6	39.0	12.85	26.25
TIP-IM$_{(ICCV'21)}$ [70]	8.0	28.2	11.6	31.2	25.2	**56.8**	34.0	51.4	19.7	41.9
Ours	**11.2**	**37.8**	**16.0**	**51.2**	**27.4**	54.0	**39.0**	**61.2**	**23.4**	**51.05**

Evaluation: Naturalness

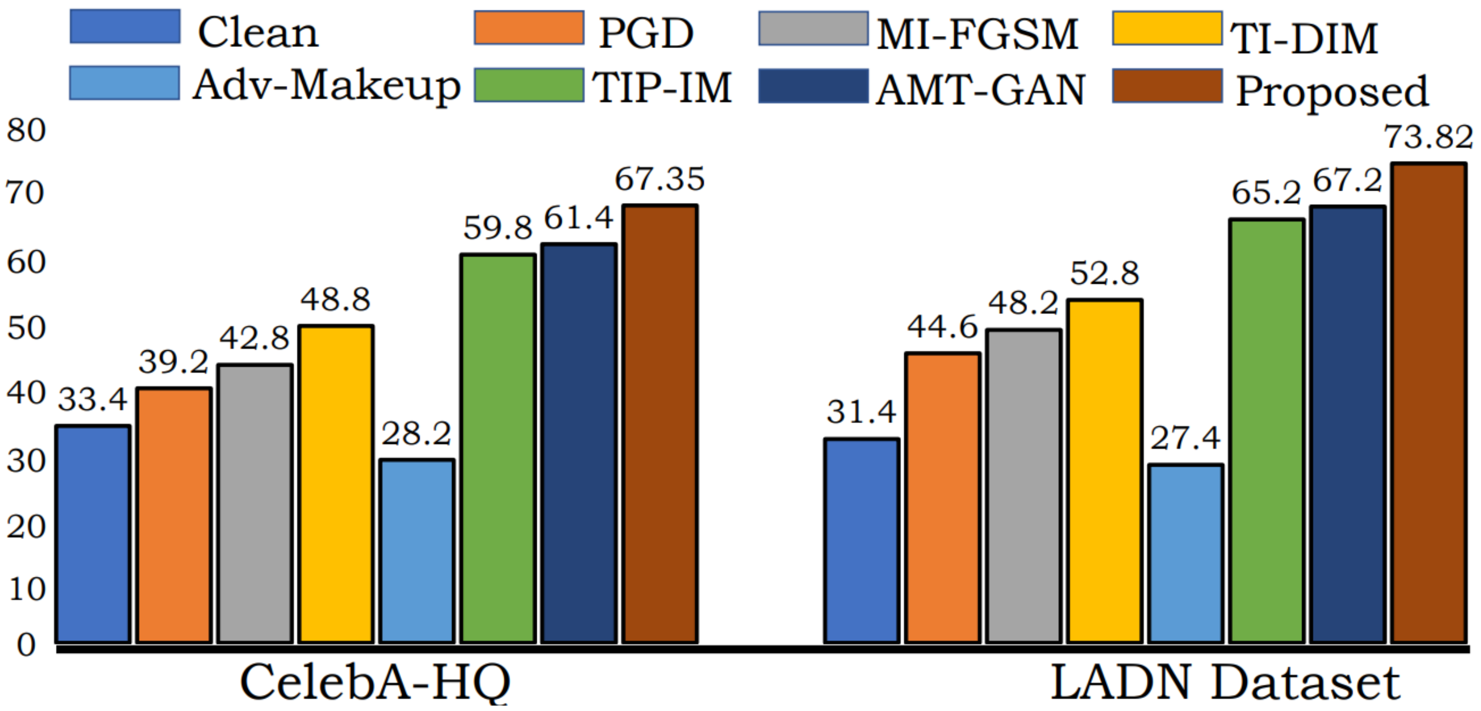
FID score to measure naturalness.



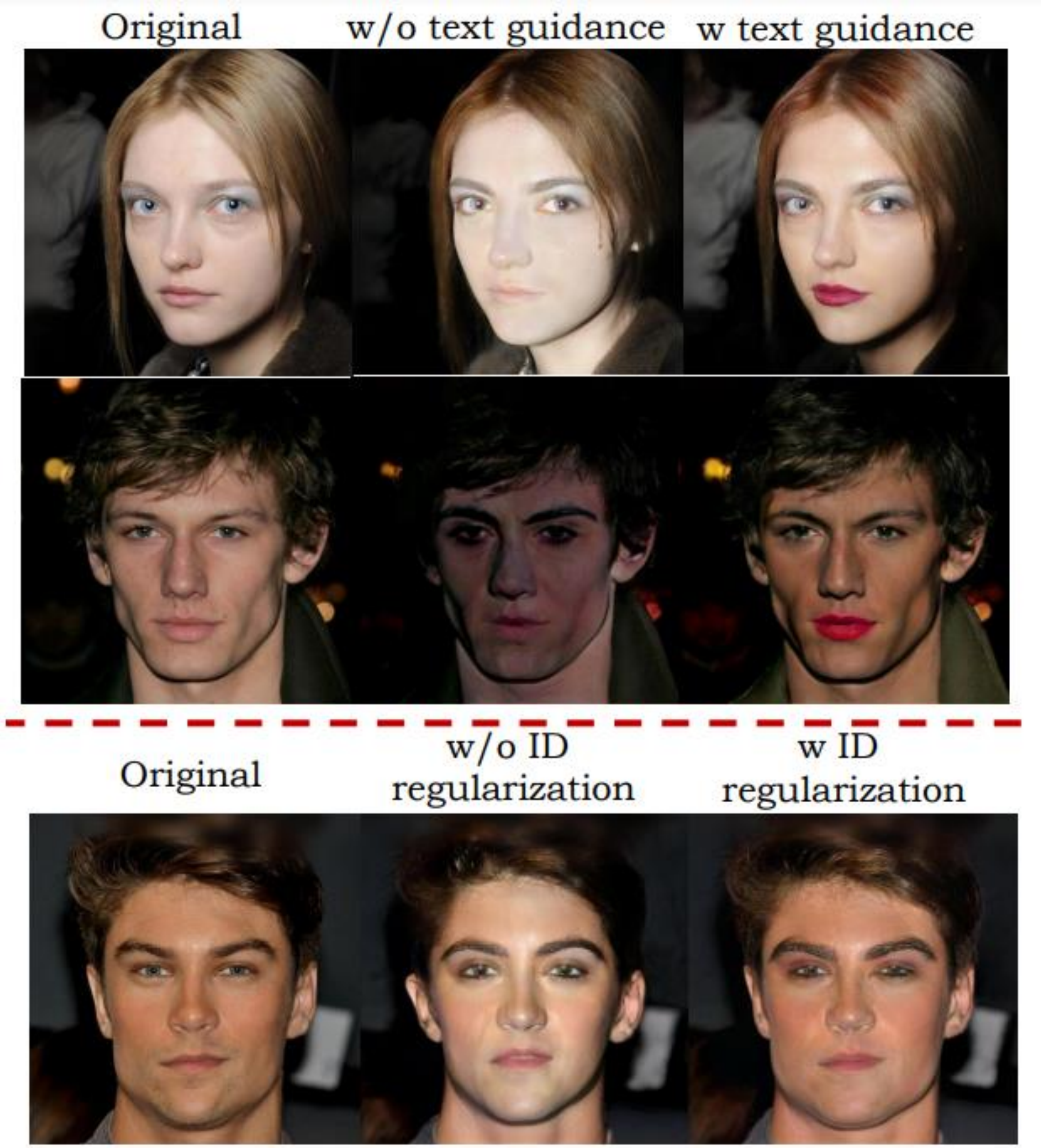
Method	FID ↓	PSR Gain ↑
Adv-Makeup [71]	4.23	0
TIP-IM [70]	38.73	35.34
AMT-GAN [22]	34.44	38.12
Ours	26.62	50.18

Evaluation: Commercial API

Average confidence score (higher is better) returned by a real-world face verification API, Face++, for impersonation attack.



Evaluation: Ablation



Thank you!