ICML
International Conference
On Machine Learning

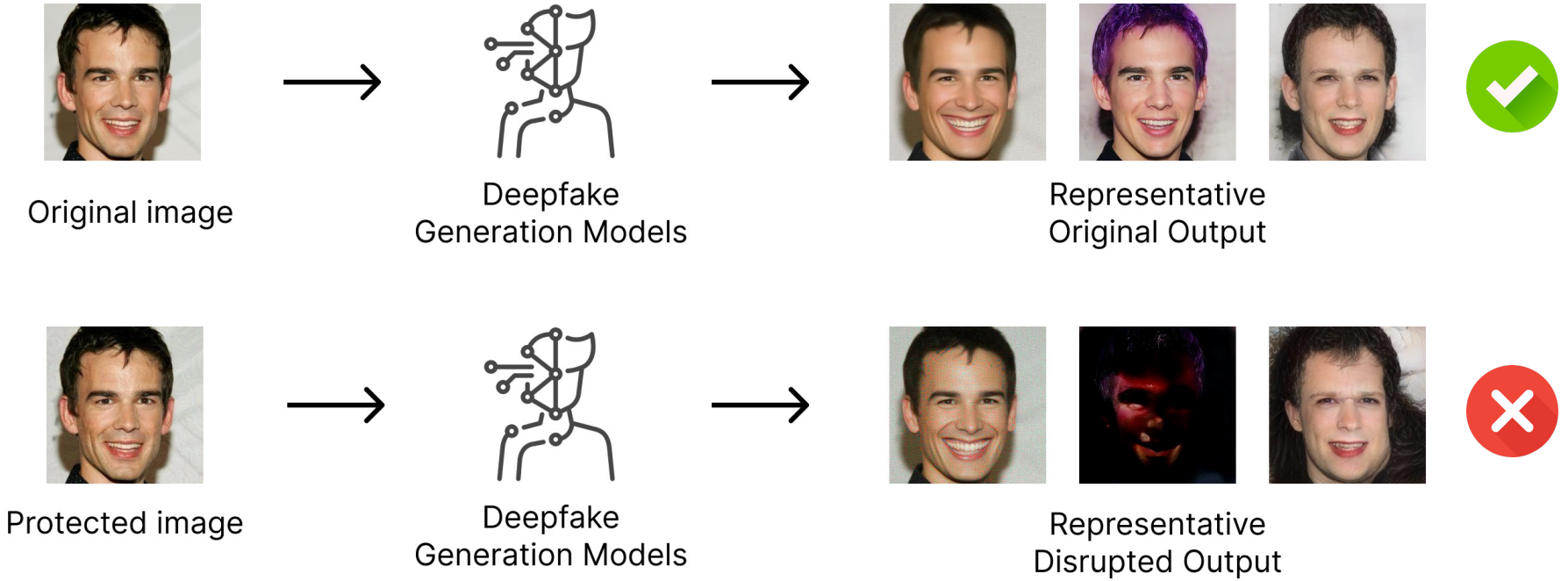
# Proactive Defense Benchmark against Deepfake Generation

Joonhyuk Baek* , Wonjune Seo* , Jae-yun Kim , Saerom Park† , Hoki Kim†


# Defend images before they're faked


Proactive defense adds an imperceptible perturbation, causing the generator to fail

# Defend images before they're faked




Proactive defense adds an imperceptible perturbation, causing the generator to fail

# Fragmented evaluation hides the real weaknesses

| Method | Pixel-wise Fidelity | Perceptual Fidelity | Visual Quality | Identity Disruption | Robustness | Transferability |
|---|---|---|---|---|---|---|
| Disrupting (Ruiz et al., 2020) | ✓ | - | - | - | ✓ | - |
| Anti-Forgery (Wang et al., 2022) | ✓ | ✓ | - | - | ✓ | ✓ |
| CMUA (Huang et al., 2022) | ✓ | - | - | - | - | ✓ |
| TCA-GAN (Dong et al., 2023) | - | ✓ | ✓ | - | - | ✓ |
| ID-Guard (Qu et al., 2025) | ✓ | - | - | ✓ | ✓ | ✓ |
| SUA (Qiao et al., 2024) | ✓ | ✓ | - | - | - | ✓ |
| FOUND (Tang et al., 2024) | ✓ | - | - | - | - | ✓ |
| Dual Defense (Zhang et al., 2024) | ✓ | ✓ | - | - | ✓ | |
| DF-RAP (Qu et al., 2024) | ✓ | ✓ | - | - | ✓ | ✓ |
| LEAT (Shim & Yoon, 2025) | ✓ | ✓ | - | ✓ | - | ✓ |
| Faceshield (Jeong et al., 2025) | ✓ | - | - | ✓ | ✓ | ✓ |
| SCOL (Lee et al., 2025) | ✓ | ✓ | - | ✓ | - | ✓ |
| NullSwap (Wang et al., 2025b) | ✓ | ✓ | - | ✓ | - | ✓ |
| FaceSwapGuard (Wang et al., 2025a) | - | ✓ | - | ✓ | ✓ | ✓ |
| **Our Benchmark** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

→

Not comparable

Blind spots stay hidden

Existing protocols can produce misleading conclusions

# Fragmented evaluation hides the real weaknesses

| Method | Pixel-wise Fidelity | Perceptual Fidelity | Visual Quality | Identity Disruption | Robustness | Transferability |
|---|---|---|---|---|---|---|
| Disrupting (Ruiz et al., 2020) | ✓ | - | - | - | ✓ | - |
| Anti-Forgery (Wang et al., 2022) | ✓ | ✓ | - | - | ✓ | ✓ |
| CMUA (Huang et al., 2022) | ✓ | - | - | - | - | ✓ |
| TCA-GAN (Dong et al., 2023) | - | ✓ | ✓ | - | - | ✓ |
| ID-Guard (Qu et al., 2025) | ✓ | - | - | ✓ | ✓ | ✓ |
| SUA (Qiao et al., 2024) | ✓ | ✓ | - | - | - | ✓ |
| FOUND (Tang et al., 2024) | ✓ | - | - | - | - | ✓ |
| Dual Defense (Zhang et al., 2024) | ✓ | ✓ | - | - | ✓ | |
| DF-RAP (Qu et al., 2024) | ✓ | ✓ | - | - | ✓ | ✓ |
| LEAT (Shim & Yoon, 2025) | ✓ | ✓ | - | ✓ | - | ✓ |
| Faceshield (Jeong et al., 2025) | ✓ | - | - | ✓ | ✓ | ✓ |
| SCOL (Lee et al., 2025) | ✓ | ✓ | - | ✓ | - | ✓ |
| NullSwap (Wang et al., 2025b) | ✓ | ✓ | - | ✓ | - | ✓ |
| FaceSwapGuard (Wang et al., 2025a) | - | ✓ | - | ✓ | ✓ | ✓ |
| **Our Benchmark** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

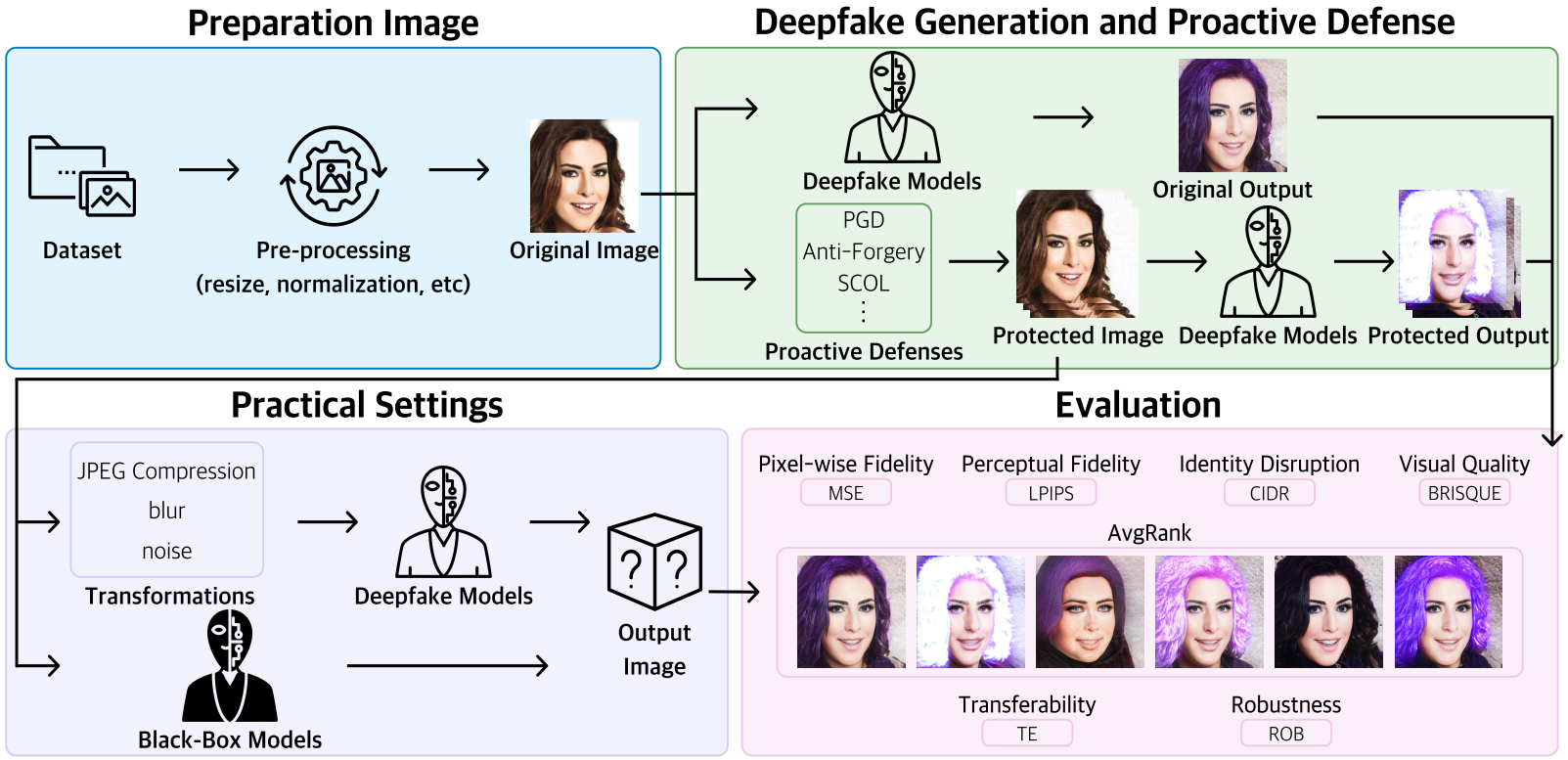
→

Not comparable

Blind spots stay hidden

Existing protocols can produce misleading conclusions

# Evaluation Framework

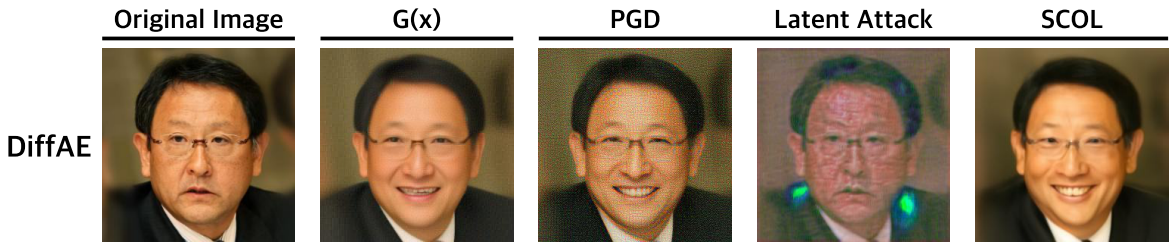



- Multi-dimensional evaluation across disruption, robustness, and transferability
- Pixel, perceptual, identity, and visual quality metrics capture distinct failure modes

# CIDR: removing generator bias

- CIDR isolates defense efficacy from generator-induced identity bias

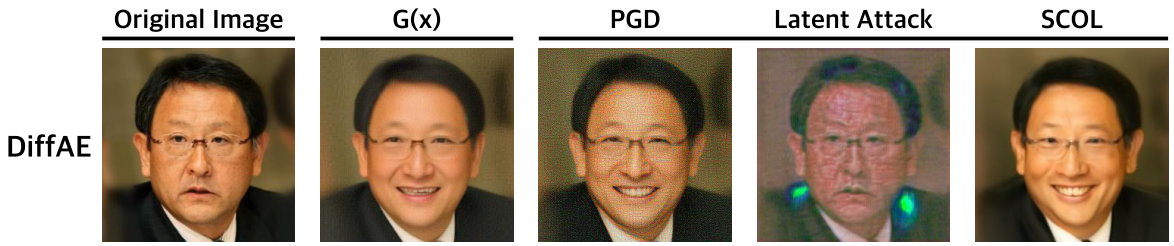
$$R_{ID}(x, x_{adv}) = 1 - \frac{L_{ID}(x, G(x))}{L_{ID}(x, G(x_{adv}))}$$



# CIDR: removing generator bias

- CIDR isolates defense efficacy from generator-induced identity bias

$$R_{ID}(x, x_{adv}) = 1 - \frac{L_{ID}(x, G(x))}{L_{ID}(x, G(x_{adv}))}$$

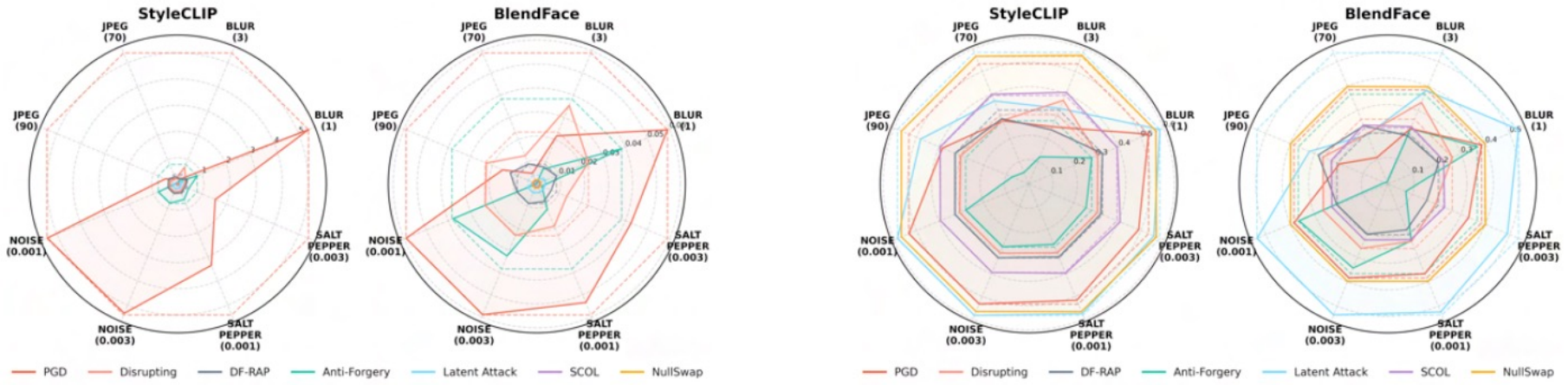



Uncalibrated identity metrics rank the wrong defense as best, CIDR is needed for a fair comparison

# Robustness vs. Disruption

- ROB measures resilience under post-processing

$$R_{ROB}(x, x_{adv}; T, L) = \frac{L(x_{ref}, G(T(x_{adv}))}{L(x_{ref}, G(x_{adv}))}$$

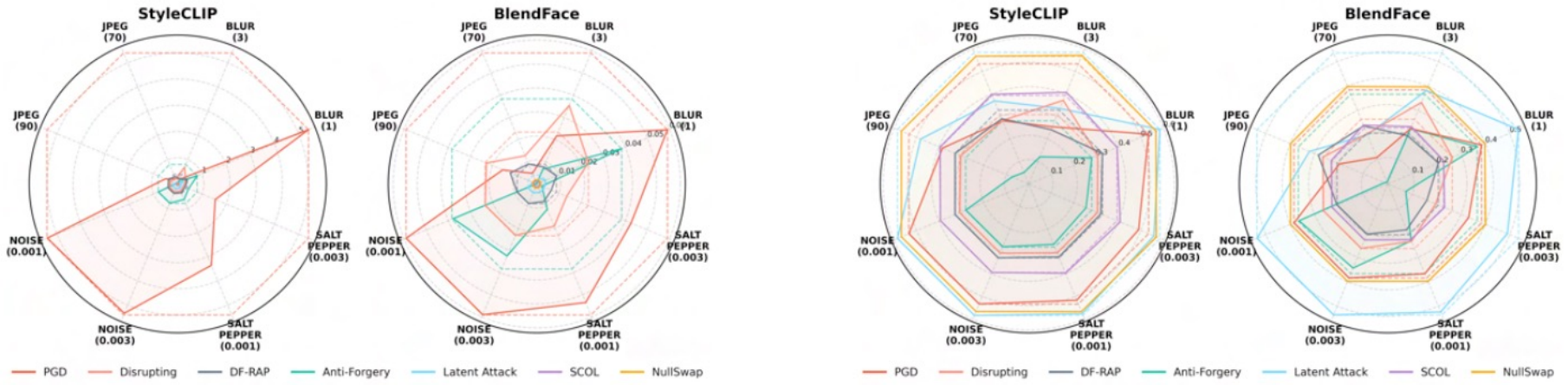



*(a)* MSE *(b)* CIDR

# Robustness vs. Disruption

- ROB measures resilience under post-processing

$$R_{ROB}(x, x_{adv}; T, L) = \frac{L(x_{ref}, G(T(x_{adv}))}{L(x_{ref}, G(x_{adv}))}$$



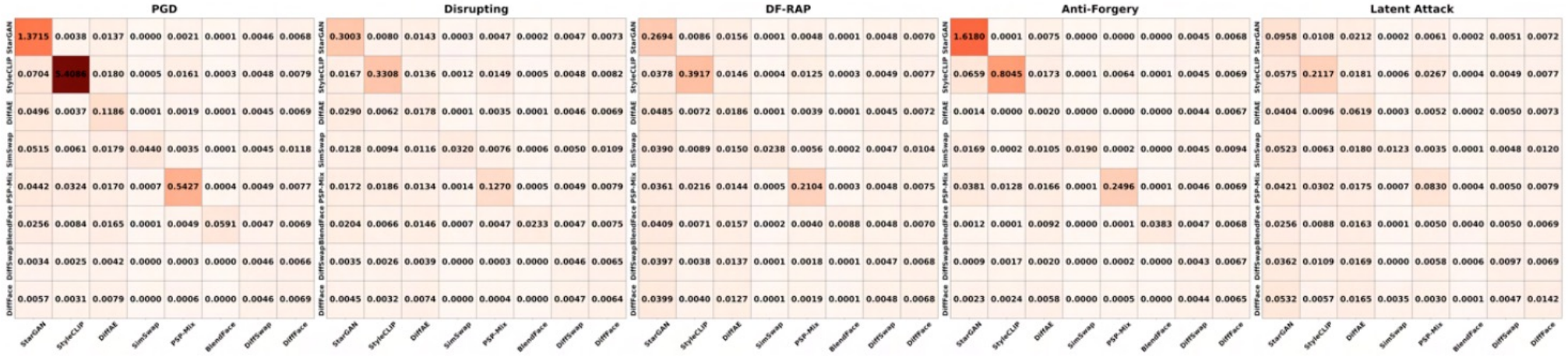

*(a)* MSE *(b)* CIDR

High white-box performance does not survive the real world, robustness must be reported alongside disruption

# Poor cross-generator transfer

- TE quantifies cross-generator generalization

$$R_{TE}^{s\to t}(x, x_{adv}^{s}, x_{adv}^{t}, L) = \frac{L(x_{ref}, G_t(x_{adv}^{s}))}{L(x_{ref}, G_t(x_{adv}^{t}))}$$

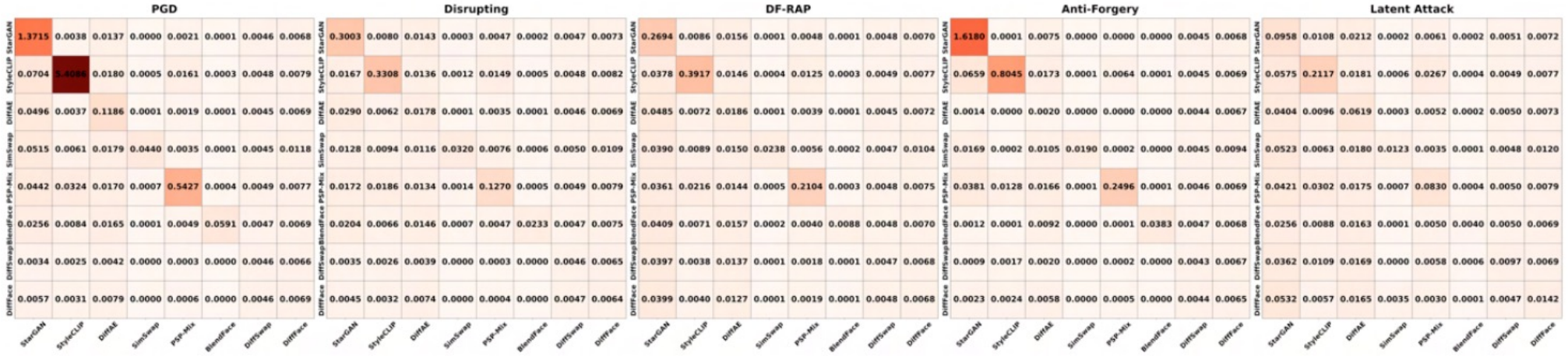



*(a)* MSE

# Poor cross-generator transfer

- TE quantifies cross-generator generalization

$$R_{TE}^{s\to t}(x, x_{adv}^{s}, x_{adv}^{t}, L) = \frac{L(x_{ref}, G_t(x_{adv}^{s}))}{L(x_{ref}, G_t(x_{adv}^{t}))}$$

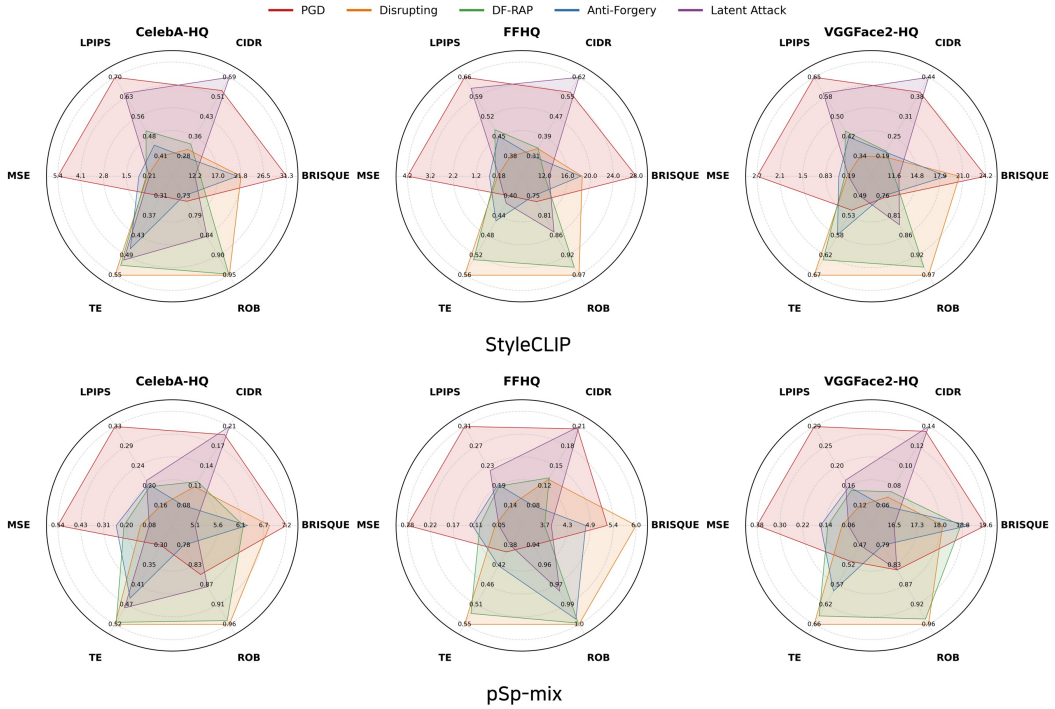



*(a)* MSE

Strong white-box results overstate protection against unseen generators, transferability must be measured

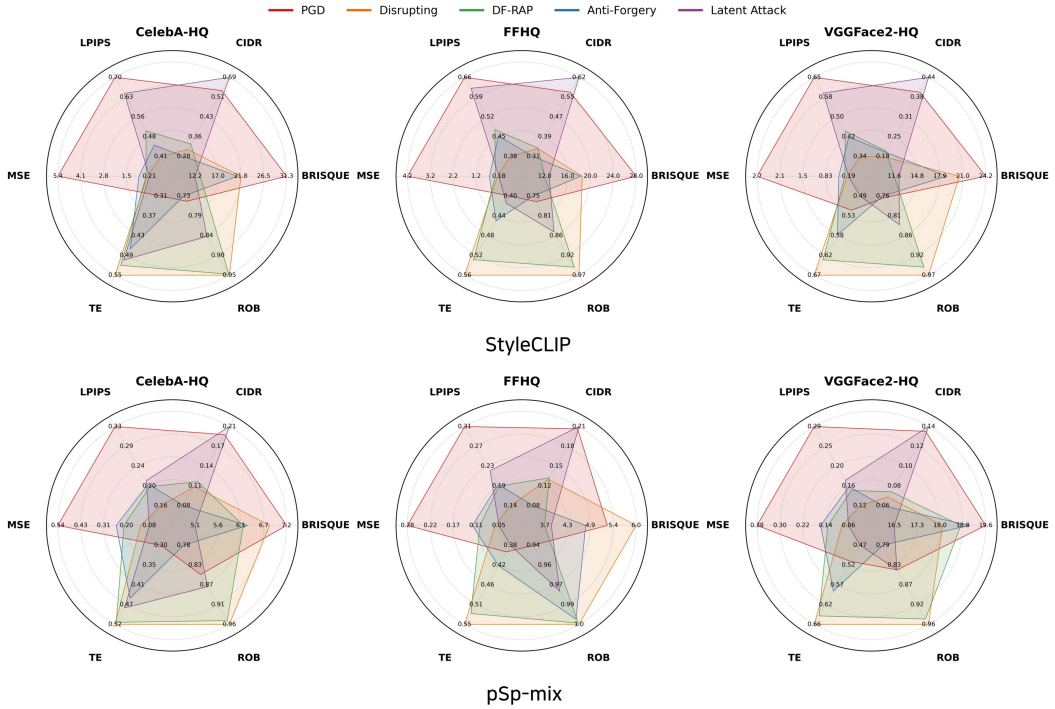
# One metric is not enough

- Single metrics overlook independent failure modes



# One metric is not enough

- Single metrics overlook independent failure modes




One metric can flip verdict, defense must be judged across complementary axes

# Conclusion

- A unified benchmark across disruption, robustness, and transferability
- Fidelity and identity metrics are orthogonal, so a single metrics can mislead
- Peak white-box performance often signals overfitting, not real protection
- CIDR corrects generator-induced identity bias for fairer evaluation

# Conclusion

- A unified benchmark across disruption, robustness, and transferability
- Fidelity and identity metrics are orthogonal, so a single metrics can mislead
- Peak white-box performance often signals overfitting, not real protection
- CIDR corrects generator-induced identity bias for fairer evaluation

→ Evaluate proactive defenses in diverse perspectives, not isolated white-box scenarios

ICML
International Conference
On Machine Learning

# Thank you!

- More details in the paper!
- Contact: jjhk7330@unist.ac.kr, wonjuneseo@unist.ac.kr

Our Project

Our Paper