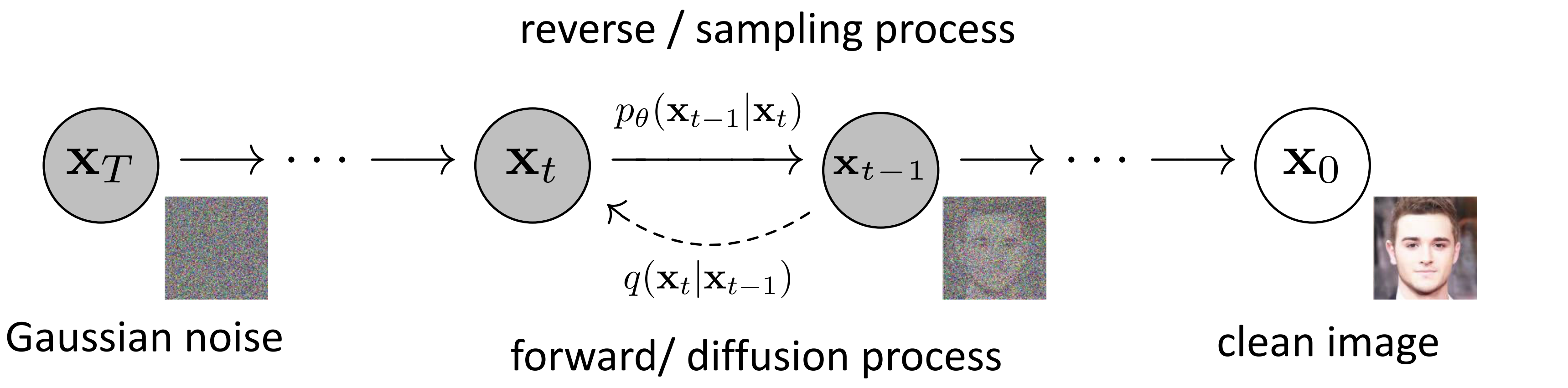
BMVC 2023

Semantic Adversarial Attacks via Diffusion Models

Chenan Wang[1] , Jinhao Duan[1], Chaowei Xiao[2], Edward Kim[1], Matthew Stamm[1], Kaidi Xu[1]
[1]Drexel University, [2]University of Wisconsin - Madison
cw3344@drexel.edu

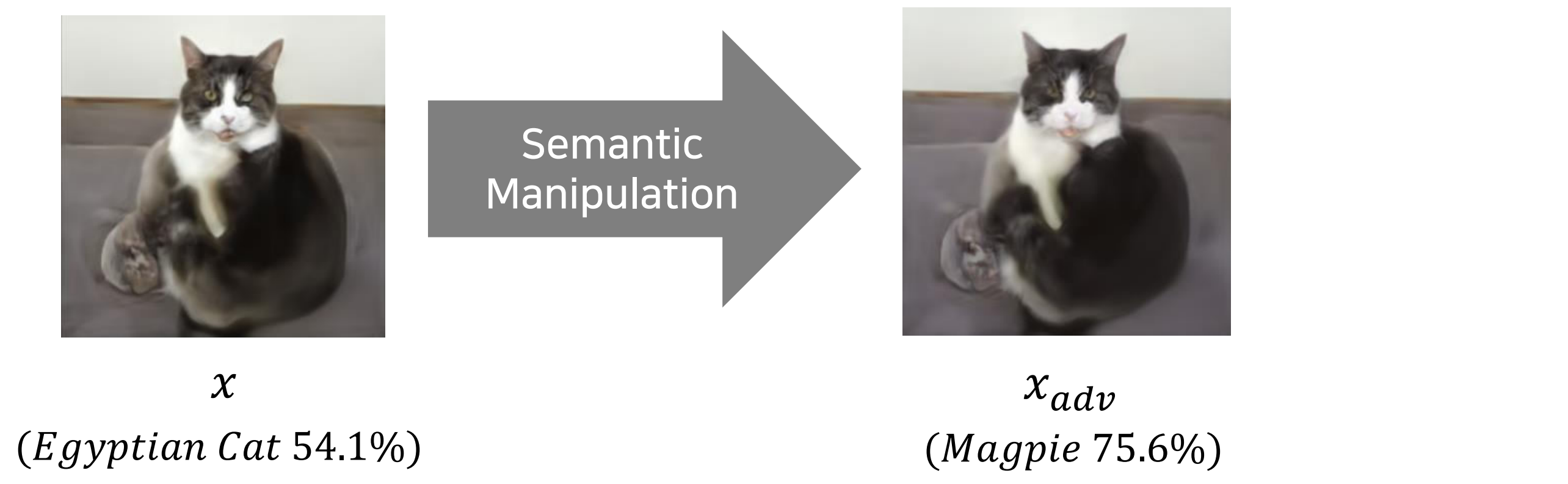
Background: Diffusion Models


reverse / sampling process
$\mathbf{x}_T$
$\mathbf{x}_t$
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$
$\mathbf{x}_{t-1}$
$\mathbf{x}_0$
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$
Gaussian noise
forward/ diffusion process
clean image

- Forward process gradually adds noise to data over time steps
- Reverse process trained to remove noise over time steps
- Sampling starts from noise and runs reverse process
- Applications includes image, audio, and text generation

The image is from Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, *33*, 6840-6851.

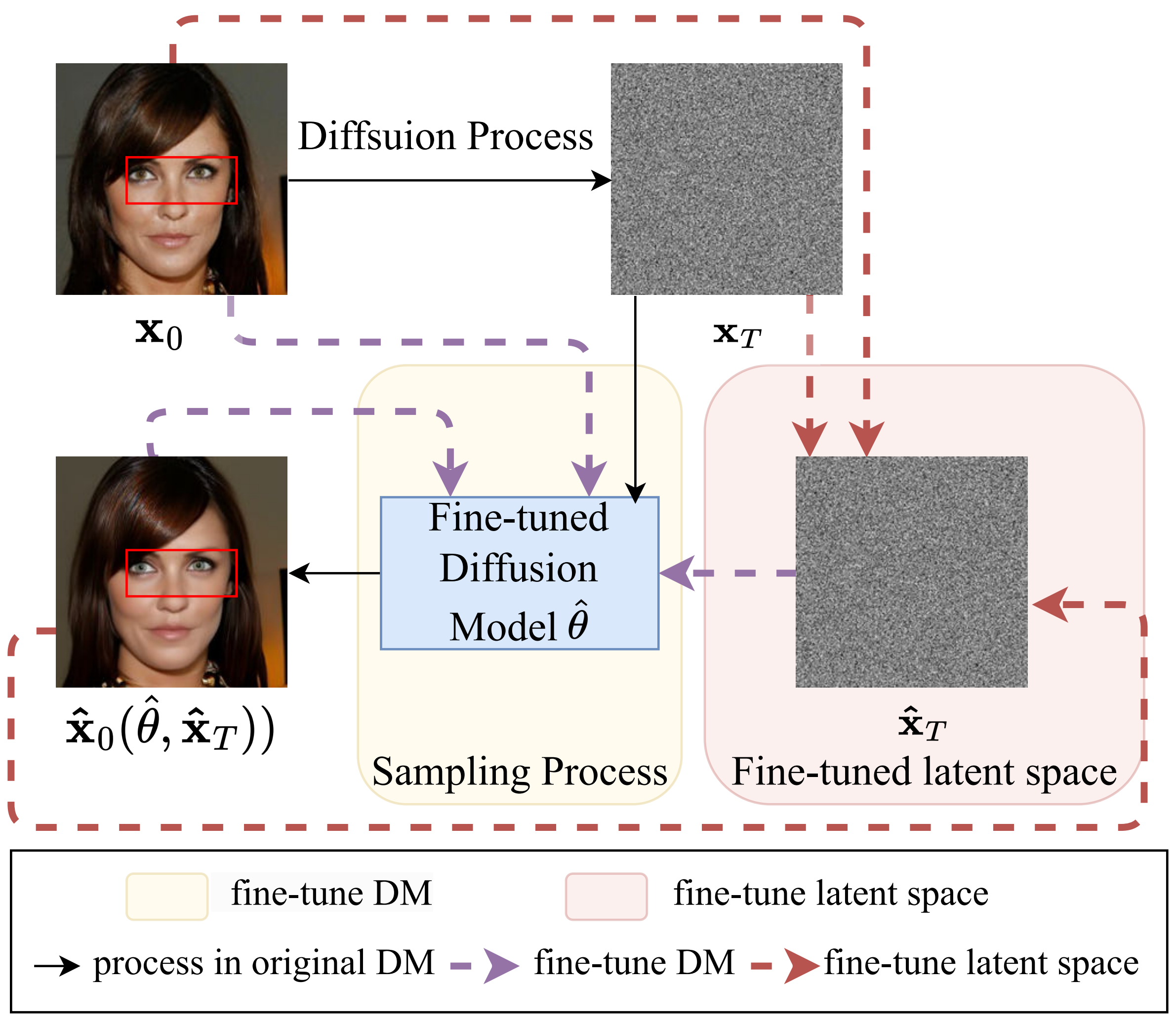
Background: Semantic Attacks



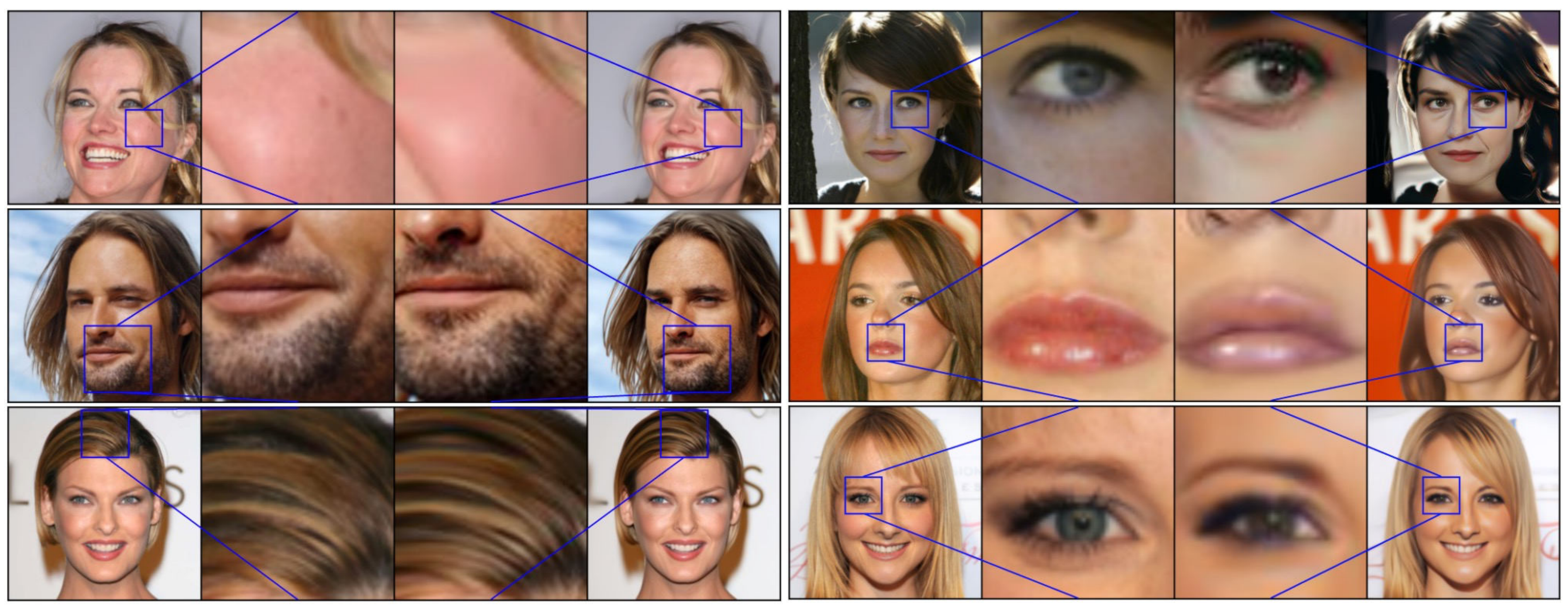
- Manipulate high-level semantic features of images, not just pixel values
- Make perceptually realistic changes to content and meaning
- Perturbations may not be norm-bounded or imperceptible
- Examples: adding/removing objects, changing color schemes, swapping backgrounds

The image is from Na, D., Ji, S., & Kim, J. (2022, October). Unrestricted Black-Box Adversarial Attack Using GAN with Limited Queries. In *European Conference on Computer Vision* (pp. 467-482). Cham: Springer Nature Switzerland.

Method: ST Approach

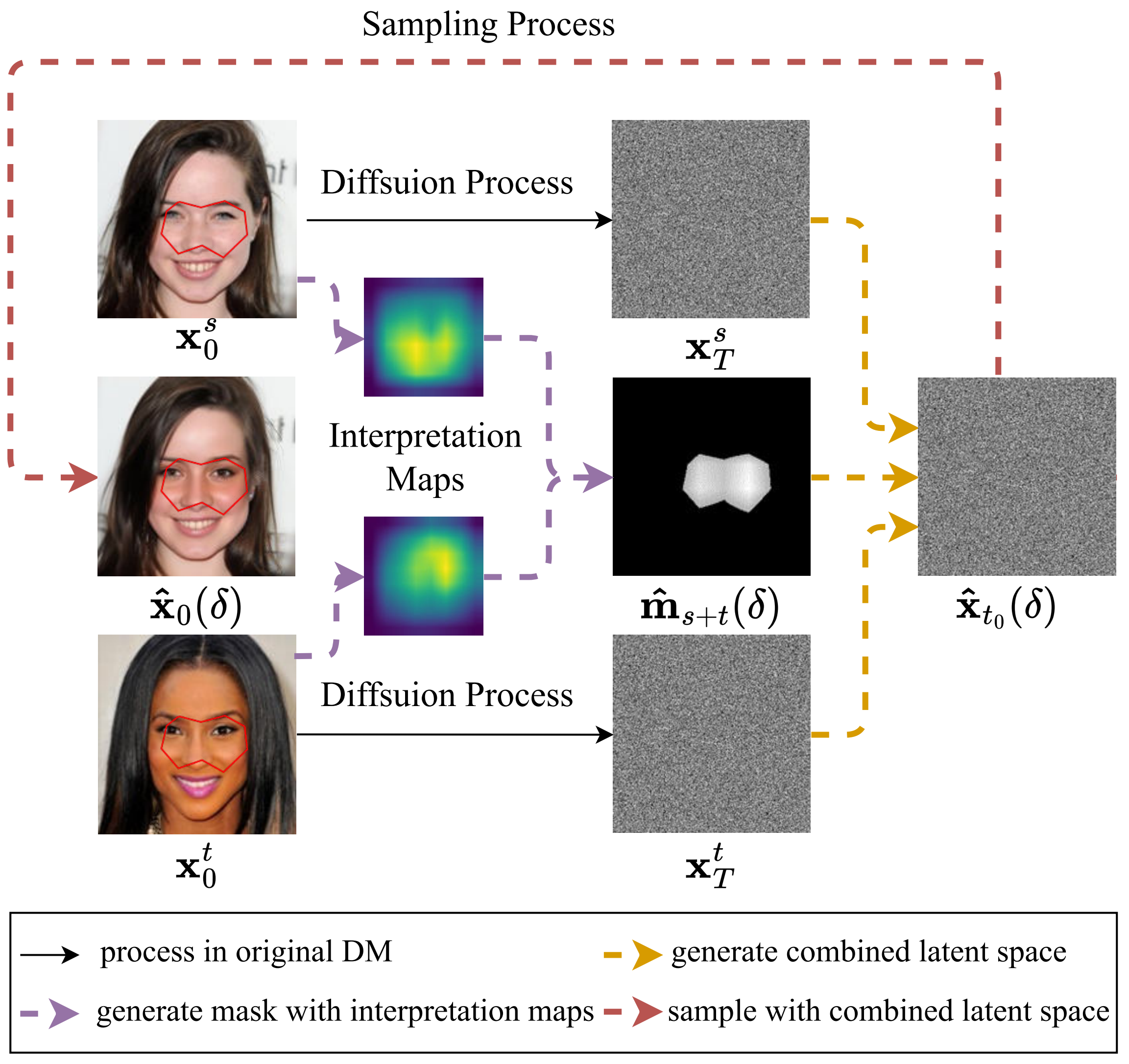


- Fine-tunes latent space and/or diffusion model parameters
- Makes minimal semantic changes to fool classifier
- Can work in white-box or black-box setting
- Achieves high attack success rate
- White-box variant has better fidelity

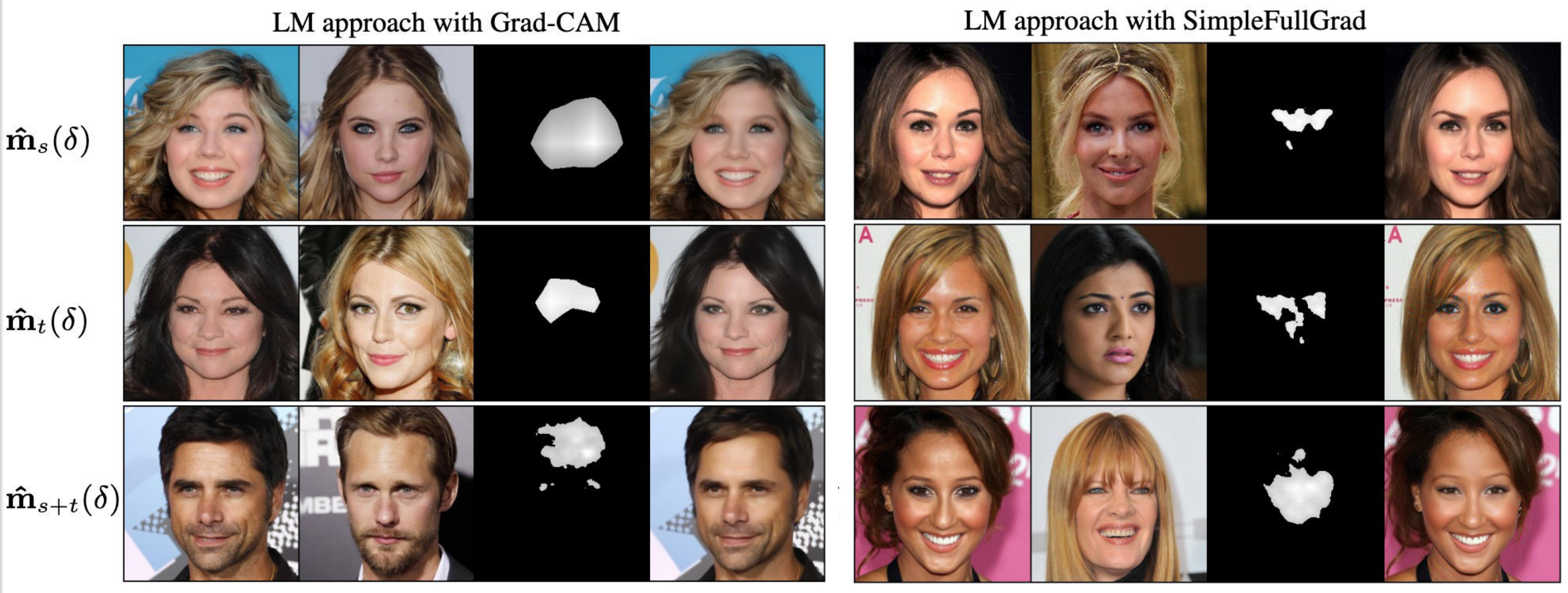


(a) white-box attack (b) black-box attack

Method: LM Approach


Sampling Process
Diffsuion Process
$\mathbf{x}_0^s$
$\mathbf{x}_T^s$
Interpretation Maps
$\hat{\mathbf{x}}_0(\delta)$
$\hat{\mathbf{m}}_{s+t}(\delta)$
$\hat{\mathbf{x}}_{t_0}(\delta)$
Diffsuion Process
$\mathbf{x}_0^t$
$\mathbf{x}_T^t$
process in original DM
generate combined latent space
generate mask with interpretation maps
sample with combined latent space

- Masks latent space with significance maps
- Transplants features from original and/or target image
- Fast method without fine-tuning diffusion model
- Achieves high attack success rate
- GradCAM gives slightly better fidelity than SimpleFullGrad
- More direct manipulation of latent space


LM approach with Grad-CAM
LM approach with SimpleFullGrad
$\hat{\mathbf{m}}_s(\delta)$
$\hat{\mathbf{m}}_t(\delta)$
$\hat{\mathbf{m}}_{s+t}(\delta)$

Experimental Results

Setting	strategy	ASR (%)↑	FID↓	KID↓	average query↓	average time (s)↓
clean images	-	-	30.67	0.000	-	-
LatentHSJA	-	100.0	83.52	0.046	1000[†]	45.87
AttAttack	-	71.80	48.92	0.018	146.82	49.71
ST approach						
fine-tune latent space	white-box	**100.0**	37.93	0.014	7.72	37.10
	black-box	59.18	114.99	0.098	43.15	206.13
fine-tune diffusion model	white-box	99.2	**36.61**	**0.006**	4.98	**30.78**
	black-box	**100.0**	96.88	0.068	11.73	66.57
fine-tune both	white-box	99.4	36.66	**0.006**	**4.96**	**30.78**
	black-box	**100.0**	94.36	0.066	11.672	64.97
LM approach						
GradCAM	$\hat{\mathbf{m}}_s(\delta)$	98.8	65.84	0.015	15.33	20.96
	$\hat{\mathbf{m}}_t(\delta)$	99.2	**64.38**	**0.014**	15.21	**18.89**
	$\hat{\mathbf{m}}_{s+t}(\delta)$	99.0	65.47	**0.014**	**14.65**	20.81
SimpleFullGrad	$\hat{\mathbf{m}}_s(\delta)$	99.6	67.10	0.016	16.17	24.03
	$\hat{\mathbf{m}}_t(\delta)$	99.6	65.21	0.016	15.32	27.48
	$\hat{\mathbf{m}}_{s+t}(\delta)$	**99.8**	65.67	0.015	14.73	23.77

[†] Elapsed time varies, depending on the query steps, which is preset by the user.

Table 1. Performance of ST and the LM approach on CelebA-HQ dataset.

- The ST approach achieves near 100% attack success rate (ASR) in all settings, with the white-box variant having better fidelity (lower FID/KID scores).
- Fine-tuning the diffusion model alone gives the best FID of 36.61 under white-box ST.
- The LM approach also gets high ASR, with GradCAM giving slightly better fidelity than saliency maps.
- Both ST and LM are much more efficient than the LatentHSJA and AttAttack baselines.

LatentHSJA: Na, D., Ji, S., & Kim, J. (2022, October). Unrestricted Black-Box Adversarial Attack Using GAN with Limited Queries. In *European Conference on Computer Vision* (pp. 467-482). Cham: Springer Nature Switzerland.
AttAttack: Joshi, A., Mukherjee, A., Sarkar, S., & Hegde, C. (2019). Semantic adversarial attacks: Parametric transformations that fool deep classifiers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4773-4783).

Experimental Results

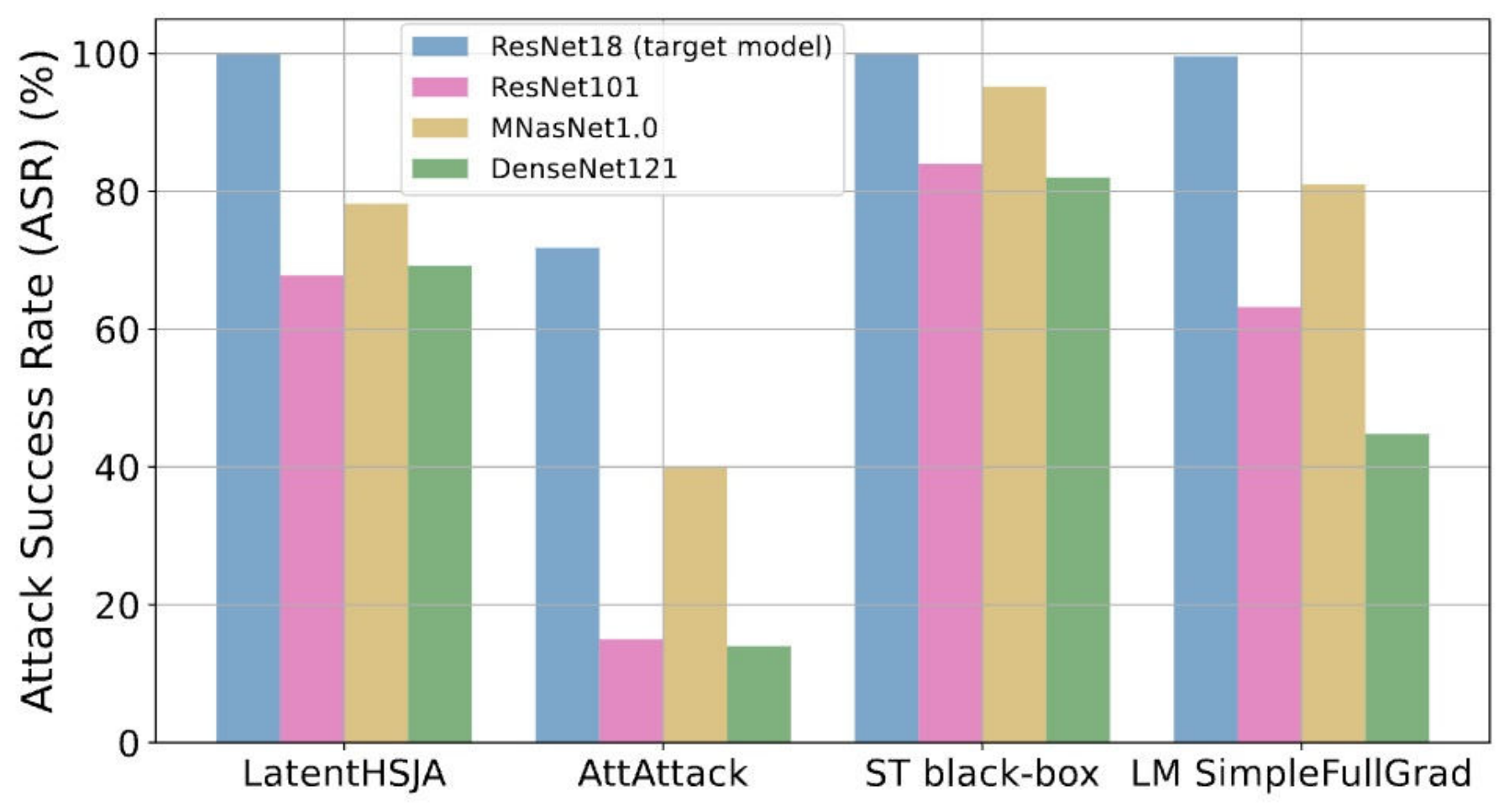


Figure 1. Transfer attack results on LatentHSJA, AttAttack, our ST and LM approach.

- We evaluate transferability of semantic adversarial attacks by generating examples to fool a ResNet18 classifier and testing them against 3 other models.
- Black-box ST approach transfers the best, maintaining high attack success rates on other models since it does not require the target model's information.
- White-box attacks tend to overfit to the target model so do not transfer as good as Black-box attacks.