THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

StyleGuard: Preventing Text-to-Image-Model-based Style Mimicry Attacks by Style Perturbations

Yanjie Li, Wenxuan Zhang, Xinqi Lyu, Yihao Liu, Bin Xiao*

Hong Kong Polytechnic University

December, 2025

*: Corresponding Author

NEURAL INFORMATION PROCESSING SYSTEMS

Research Background

- Diffusion models have been used for **style mimicry** (generate images in a specific artist's artistic style) through methods like DreamBooth [1] . This has raised concerns about intellectual property protection.
- For example, an attacker can fine-tune a stable diffusion model using a small number of Van Gogh paintings on text ”<Van Gogh>”, and then prompt this new model to “generate a painting of trees in the style of <Van Gogh>."

Example images generated by prompt: generate a painting of trees in the style of <Van Gogh>.

[1] Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023.

Problem Statement

➢ Objective:

- The goal is to create a protected image set X_p^* with optimized noises. Fine-tuning on this image set using DreamBooth method will lead to a poorly performing Stable Diffusion model, resulting in generated images (X_{gen}) that exhibit styles differing significantly from those in the original image set (X_c).
- This objective can be expressed as a bilevel optimization problem:

$$X_p^* \in \arg\max_{X_p, \theta^*} L_{\text{dis}}(X_{gen}; X_c)$$

$$\text{where} \quad \theta^* \in \arg\min_{\theta} \{L_{\text{gen}}(\mathbb{T}(X_p); c, \theta)\}.$$

- **c**: Class-wise conditional vector.
- X_{gen}: Generated images from the fine-tuned LDM model
- L_{dis}: Perception-aligned distance function that evaluates style discrepancy between generated and reference images .
- L_{gen}: Fine-tuning loss function (such as a Dreambooth-like loss function).

Exist Defense Strategies

- To address the unauthorized style mimicry issue, perturbation-based methods have been developed, which add subtle image perturbations to the unprotected images to disrupt generative models.
  - **Glaze**: Minimizes the feature distance between perturbed and target images while preserving perceptual similarity.
  - **AdvDM**: Disrupts the denoising process in pre-trained diffusion models, reducing the likelihood of generating perturbed images.
  - **Mist**: Integrates semantic and textual loss for enhanced protection, although it may result in unnatural textures.
  - **Anti-DreamBooth**: Alternately updates the diffusion model and protected images to defend against personalization attacks.
  - **MetaCloak**: Builds upon Anti-DreamBooth by incorporating transformations (e.g., Gaussian blur, cropping) to enhance robustness.

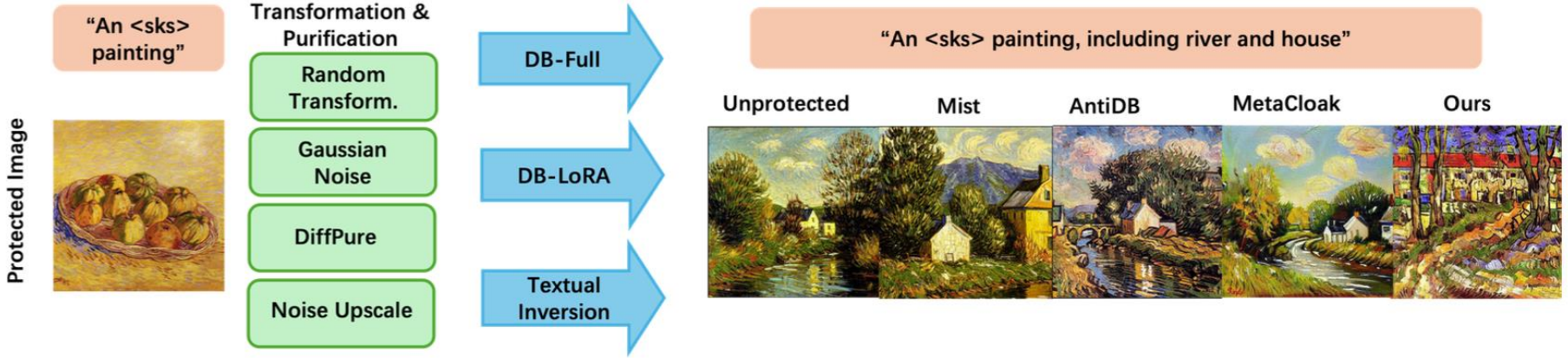
Challenges in Exist Protection

- **Current Methods are challenged by new techniques**
  - While existing perturbation-based methods show robustness to simple transformations, they are increasingly challenged by new techniques.
  - **Purification-Based Methods:**
    - DiffPure & Noise Upscaling (ICLR'25):
    - **Process**:
      - Add noise to images
      - Use Latent Diffusion Models (LDMs) as purifiers or upscalers to remove noise
    - Proven to effectively eliminate protective noise
    - Render many existing protective methods ineffective
- **Present methods show limited cross-model transferability.**
  - Different Diffusion Model Architectures
  - Different finetuning methods, such as DreamBooth, LoRA and textual inversion.

Overview of StyleGuard

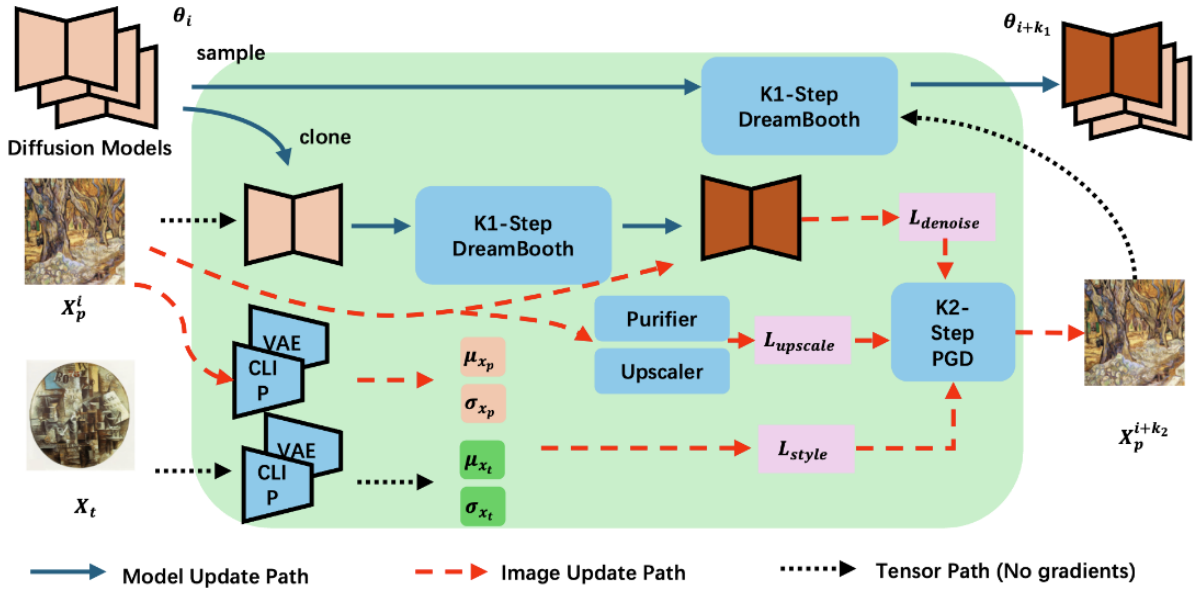
➢ Comparison of Defensive Performance against **Purification Transformations**

- **Previous Methods**:
  - **Mist**: Ineffective against purification attacks.
  - **Anti-DreamBooth**: Fails to defend against DiffPure and Noise Upscaling.
  - **MetaCloak**: Also unable to withstand these transformations.
- **Our Proposed Method**
  - Successfully resists style mimicry attacks.
  - Maintains defense under various transformations and customization methods.



Methodology

- We alternatively update the diffusion model and the protected images.
- This can optimize the perturbation noise to disrupt a dynamic, unknown generator that is finetuned from the perturbed images themselves.
- Ensemble image encoders and purifiers are included to compute the **style loss** and **upscale loss** to improve cross-model transferability and the robustness to purifications.



Loss Functions

- Style Loss
  - **Objective**: Maximize the **style-related feature** distance between reference and perturbed images, aligning it with target images.
  - **Definition:**

$$L_{\text{style}} = \mathbb{E}_{f\sim F}\mathbb{E}_{x_p\sim X_p, x_t\sim X_t}\left(\left\|\mu_{x_p}-\mu_{x_t}\right\|_2^2-\left\|\sigma_{x_p}-\sigma_{x_t}\right\|_2^2-\left\|\mu_{x_p}-\mu_{x_c}\right\|_2^2+\left\|\sigma_{x_p}-\sigma_{x_c}\right\|_2^2\right)$$

  - **Variables:**
    - μ and σ: Mean and variance of latent features encoded by the VAE or CLIP encoder.
    - X_t: Set of target images with distinct styles from the reference images X_c.
  - By disturbing style-related features, StyleGuard complicates the connection between style features and identifiers, enhancing protection against attacks like DreamBooth or Textual Inversion.

Loss Functions

- Denoising Error Loss Function
  - **Objective**: : This loss function aims to ensure that the perturbations are effective against the stable diffusion model, maintaining robustness through the optimization process.
  - **Definition**:

$$L_{\text{denoise}} = -\mathbb{E}_{\theta\sim\Theta}\mathbb{E}_{x_{p,0},t,c,\epsilon\sim\mathcal{N}(0,1)} \|\epsilon - \epsilon_\theta(x_{p,t+1}, t, c)\|_2^2$$

  - **Variables**:
    - ϵ: Noise added to the image.
    - $x_{p,t+1}$: Perturbed image at the next timestep.

  - This is equivalent to maximizing the Dreambooth training loss function.

Loss Functions

- Upscale Loss Function
  - **Objective**: Encourage purifiers and upscalers to amplify protective perturbations rather than diminish them.
  - **Definition**:

$$L_{\text{upscale}} = -\mathbb{E}_{\theta_{\mathbb{T}} \sim \Theta_{\mathbb{T}}} \mathbb{E}_{x'_{p,0}, t, c, \epsilon \sim \mathcal{N}(0,1)} \left\| \epsilon - \epsilon_{\theta_{\mathbb{T}}}(x'_{p,t+1}, t, c) \right\|_2^2, \text{ where } x'_p = x_p + \delta N(0,1),$$

  - **Variables**:
    - $x' = x + \delta N(0,1)$:Perturbed image with added noise.
    - Θ_T: Parameters of the purification and upscaling models.
  - By maximizing the denoising error across various models, this function strengthens the defense against noise purifiers, making it more challenging for attackers to diminish the protective features.
- Total Loss Function:

$$L_{\text{styleguard}} = L_{\text{denoise}} + \eta L_{\text{upscale}} + \lambda L_{\text{style}},$$

StyleGuard Algorithm

- Generating Perturbation Noises
  - A straightforward approach to tackle the bilevel problem in slide 3 is to unroll all training steps and optimize the images through backpropagation. **However**, this would cause a very large computation graph that would exceed the capacity of most current machines.
  - To solve this, we iteratively update the stable diffusion model and the image.
  - First, we update stable diffusion model for K1 steps:

$$\theta'_{t,j+1} = \theta'_{t,j} - \beta \nabla_{\theta'_{t,j}} L_{\text{gen}}(X_p^t; \theta'_{t,j}),$$

  - To improve the robustness to other transformations like crop and resize, we involve random transformations T' in the optimization process.
  - Then we update the perturbed image X_p using PGD with the attack budget B_∞ with K2 steps:

$$X_p^{t+1} = \mathbb{E}_{g \sim \mathbb{T}'} \left[\Pi_{B_\infty} \left(X_p^t + \alpha \, \text{sign} \left(\nabla_{X_p^t} L_{\text{styleguard}} \right) \right) \right],$$

  - Repeat this process until the maximum number of iterations.

Experiment Setup

Datasets

- **WikiArt**:
  - 42,129 artworks from 195 artists (e.g., impressionism, cubism).
  - Style mimicry attacks: Randomly selected 40 artists, 20 artworks each (10 for training, 10 for evaluation).
- **CelebA**:
  - 100 identities selected, 10 images per identity for training and evaluation.

Implementation Details

- **Models**: SD v1.4 and SD v1.5 used for perturbation.
- **Attack budget**: Set to 8 out of 256, consistent with baselines.
- **Encoders for style loss**: VAE, OpenCLIP-ViT-H-14, OpenCLIP-ViT-bigG-14.
- **Fine-tuning methods evaluated**: DreamBooth (Full-FT, LoRA-FT) and Textual Inversion.

Experiment Setup

- **Preprocessing Settings**:
  - **Random Transformations**: Gaussian noise, center cropping, resizing.
  - **DiffPure**: Utilized Guided Diffusion Model and DDPM during training.
  - **Noise Upscaling**: Perturbations generated on SD-x4-upscaler; evaluated using SD-x2-latent-upscaler.

➢ **Baseline Settings**

- Compared against: Glaze, Mist, Anti-DreamBooth, MetaCloak, SimAC.
- Attack budgets set to 8/256 (except Glaze, which uses LPIPS as constraint).

➢ **Evaluation Metrics**

- **Style Mimicry**: Fréchet Inception Distance (FID), precision, mimicry success rate (human-annotated).
- **Personalization Attacks**: Identity Match Score (IMS) to assess semantic closeness.

Experiment Results

Table 1: Comprehensive evaluation of text-to-image protection methods under different transformations for DreamBooth on the WikiArt Dataset. Metrics reported are FID↑ (higher better) and Precision↓ (lower better). The best data are shown in bold, and the second runners are in gray.

Method	No Prep.		Crop+Resize		Gauss. Noise		DiffPure		Noise Up.	
	FID	Prec.	FID	Prec.	FID	Prec.	FID	Prec.	FID	Prec.
No Protect	233.78	0.60	275.41	0.65	238.25	0.62	237.89	0.68	236.58	0.60
Glaze	333.89	0.15	315.22	0.40	340.10	0.35	318.94	0.30	312.73	0.60
Mist	382.50	**0.00**	295.28	0.46	275.77	0.48	290.45	0.42	256.65	0.45
AntiDB	327.01	0.05	310.88	0.25	322.15	0.20	305.74	0.35	293.14	0.50
MetaCloak	382.00	0.05	362.52	0.20	355.26	0.18	318.87	0.25	295.20	0.40
SimAC	407.40	**0.00**	365.47	0.06	380.45	0.04	290.15	0.38	284.52	0.45
L_{denoise}	348.15	0.03	355.77	0.30	362.42	0.15	358.89	0.12	310.54	0.20
$L_{\text{denoise+style}}$	389.33	0.01	382.45	0.25	380.21	0.08	385.77	0.06	375.92	0.10
StyleGuard	**428.70**	**0.00**	**405.31**	**0.05**	**420.74**	**0.02**	**418.33**	**0.03**	**401.80**	**0.00**

Experiment Results

- **Key Findings**:

- **Baseline Reference**:
  - **Mist**: Shows weaknesses to transformations (cropping, resizing, Gaussian noise) with Precision scores of 0.46 and 0.48.
  - **MetaCloak & SimAC**: More robust to simple transformations but falter against purifiers/upscalers, yielding low Precision (MetaCloak: 0.20; SimAC: 0.06).

- **StyleGuard Performance**:
  - Outperforms all existing techniques in both scenarios.
  - Achieves the highest FID and lowest Precision across all transformations, demonstrating strong protective efficacy, especially against purifiers/upscalers.

Experiment Results: Visualization

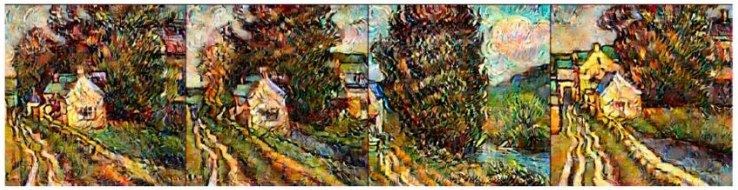
Style Mimicry on Clean Images

Style Mimicry on Images Protected by StyleGuard

Experiment Results

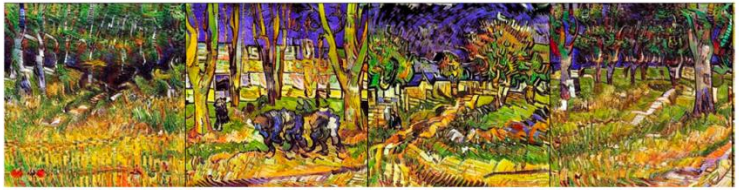
- Ablation Study

(1) Train on clean images



(2) Style Loss + Denoise Loss, without Noise Upscale

(3) Style Loss+Denoise Loss, with Noise Upscale



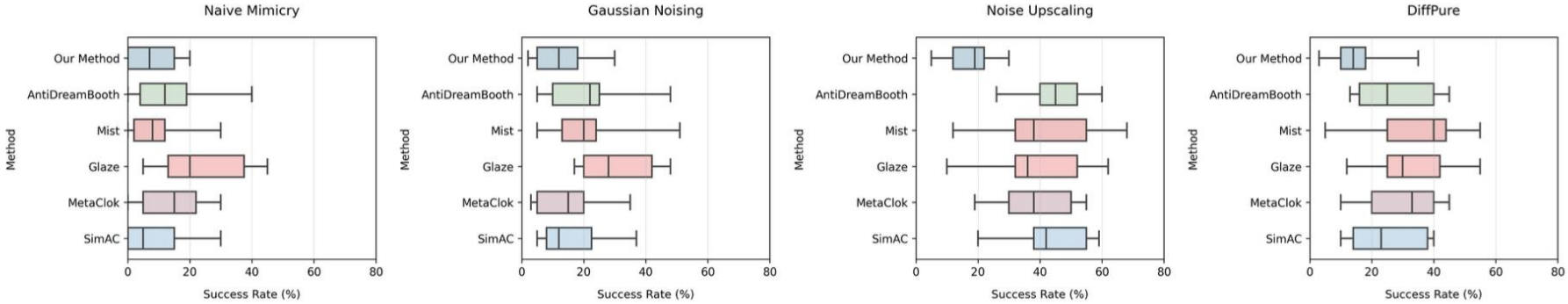
(4) Style Loss+Denoise Loss+Upscale Loss, with Noise Upscale

Figure 3: Visualizing the effects of different loss functions. It is shown that only using denoise loss and style loss cannot defend the Noise Upscale well, as shown in (3). With the Upscale Loss, the image quality significantly decreases even with Noise Upscale, as shown in (4).

Experiment Results

➢ Human Evaluation

- We asked users to compare generated images based on clean and protected training images using the question: "**Based on the image style and quality, which image better fits the reference samples?**"
- Our method achieves lower mimicry success rate, indicating stronger perturbation noises affecting the image quality.



Experiment Results

- Cross-model Transferability
  - **Table 2**: Shows the ratio of FID scores (%) for images generated on evaluation models versus substitute models.
  - **Key Finding**: StyleGuard demonstrates better transferability across different SD models compared to Anti-DreamBooth.
- Reason for Improved Transferability
  - Incorporates **style loss** alongside denoising loss.
  - Perturbs global style-related features that remain independent of model parameters, enhancing robustness.

Table 2: Cross-model transferabilities for AntiDreamBooth and StyleGuard.

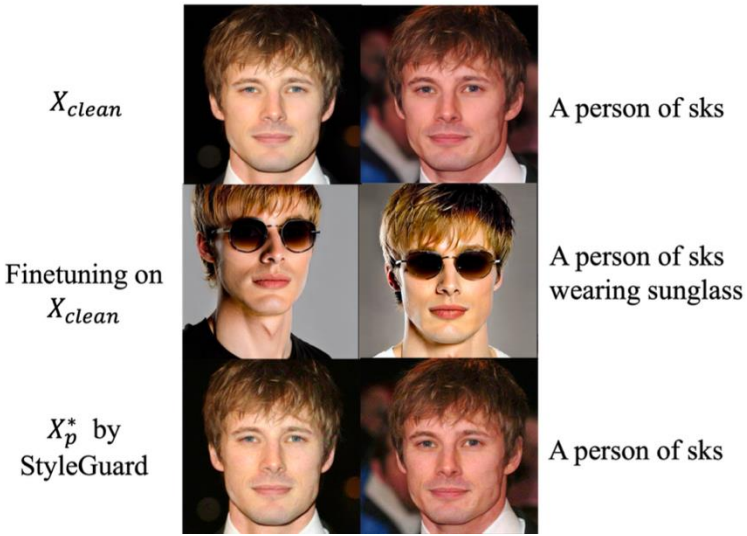
Surrogate ↓	Evaluation model		
	SD v1.4	**SD v1.5**	**SD v2.1**
SD v1.4	100.0/100.0	85.5/96.5	76.5/92.4
SD v1.5	84.8/96.2	100.0/100.0	73.5/92.5
SD v2.1	73.5/89.2	76.4/92.5	100.0/100.0

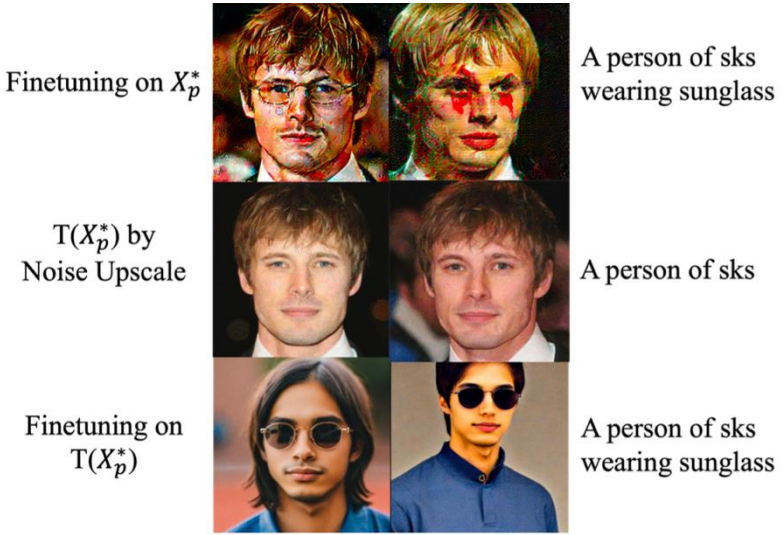
Table 3: Transferability to different fine-tuning methods.

Finetuning Method	LoRA/SD v2-1		LoRA/SD XL	
	FID ↓	Prec. ↑	FID ↓	Prec. ↑
Clean (Baseline)	210.45	0.75	215.60	0.72
AntiDreamBooth	285.20	0.45	365.80	0.28
MetaCloak	320.85	0.28	435.60	0.22
SimAC	375.90	0.20	445.75	0.12
Ours	366.78	0.16	464.45	0.00

Experiment Results on Personalization Attacks

- StyleGuard also successfully defends against DreamBooth personalization attacks, even with Noise Scaling preprocessing applied to face images.





Conclusion

- We propose StyleGuard, a robust method designed to effectively protect artists from DreamBooth-based style mimicry. StyleGuard accounts for various preprocessing techniques that attackers may employ, enhancing its practical effectiveness.
- StyleGuard demonstrates strong robustness against various purification methods by including an upscale loss that maximizes denoise-error loss across ensemble purifiers and upscalers.
- StyleGuard improves cross-model transferability by introducing a style loss that aligns the style-related features of protected images with those of target images.
- Experiments on the WikiArt and CelebA datasets demonstrate that StyleGuard offers enhanced protection against style mimicry and identity customization.
- Compared to previous methods, our approach shows improved efficacy and practical effectiveness by considering various preprocessing techniques and model-agnostic scenarios.