B-Free

A Bias-Free Training Paradigm for More General AI-Generated Image Detection

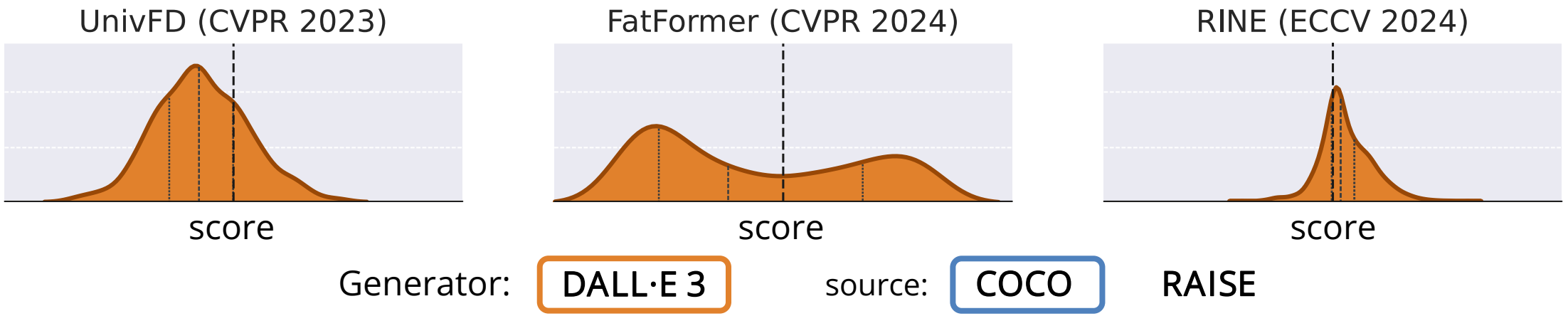
Fabrizio Guillaro*, Giada Zingarini*, Ben Usman**,
Avneesh Sud**, Davide Cozzolino*, Luisa Verdoliva*

* University of Naples Federico II
** Google DeepMind

CVPR Nashville JUNE 11-15, 2025
Sat 14 Jun, 3 PM - #277

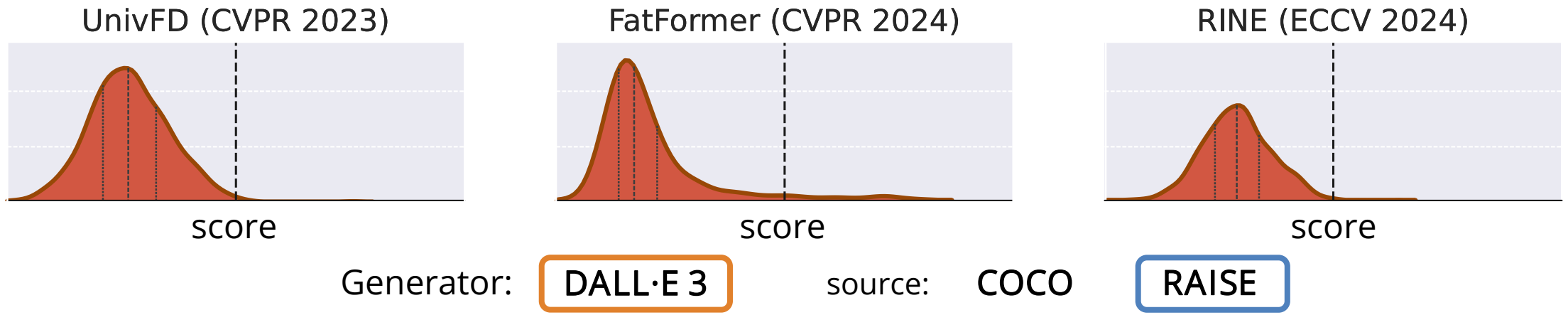
Problem

- Current synthetic image detectors perform well in controlled scenarios, but **struggle to transfer** to real-world applications
- We believe that this is caused by the **poor quality of training data**, which is often influenced by several **biases**
- This leads to inconsistent performance even for synthetic images generated by the same model



Problem

- Current synthetic image detectors perform well in controlled scenarios, but **struggle to transfer** to real-world applications
- We believe that this is caused by the **poor quality of training data**, which is often influenced by several **biases**
- This leads to inconsistent performance even for synthetic images generated by the same model



Our work

- **A well-designed forensic detector should detect generator specific artifacts rather than reflect data biases**
- While most research focuses on developing **complex algorithms**, we shift the focus on improving the training paradigm
- We propose **B-Free**: a new training paradigm to prevent biases
- The curated **training dataset** and the **designed augmentation** proposed help to obtain calibrated results in real-world scenario

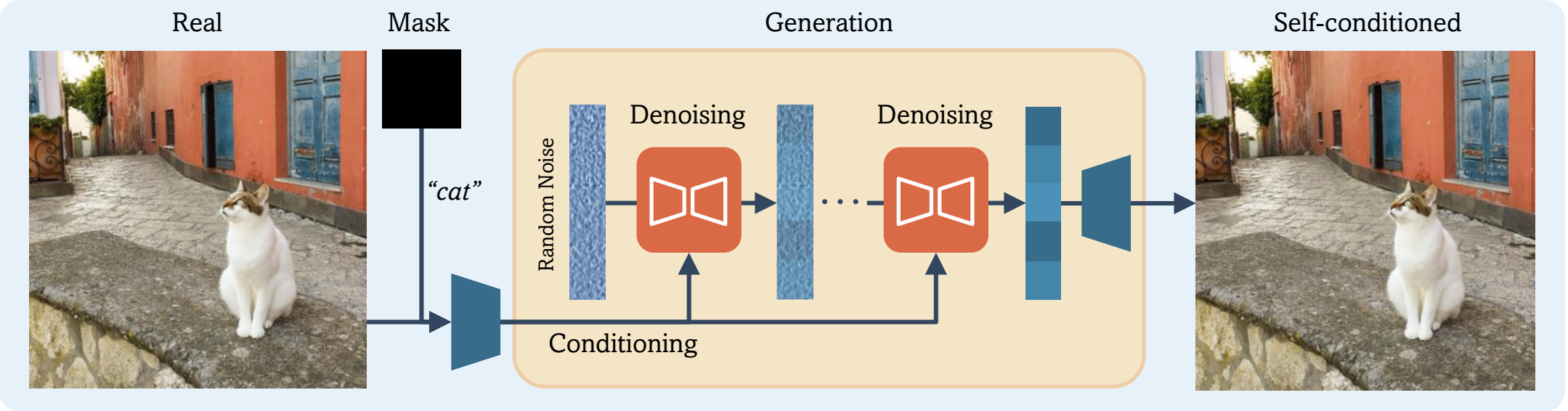
Proposed approach

- Our new training paradigm relies on the use of:

Real

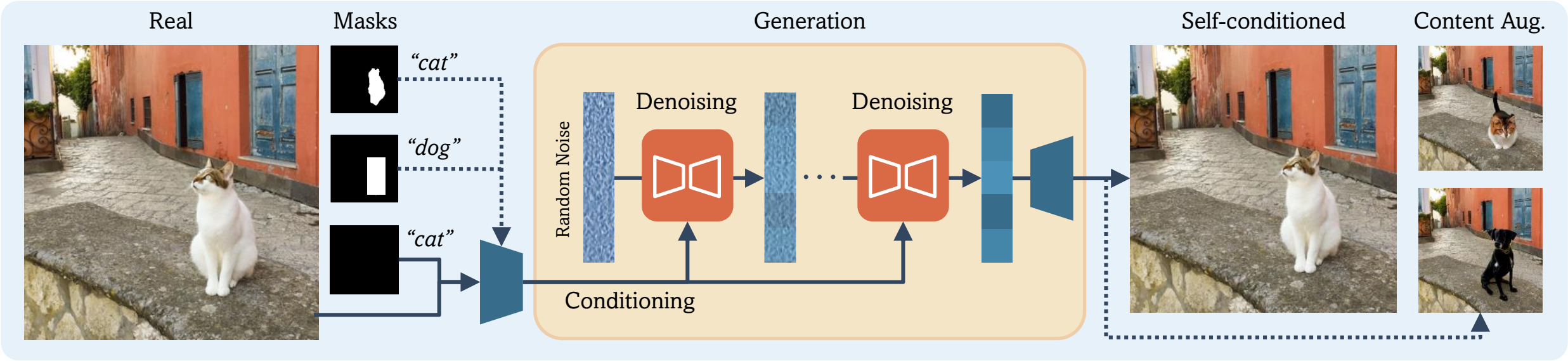
Proposed approach

- Our new training paradigm relies on the use of:
  - **self-conditioned reconstructions** of real images using Stable Diffusion 2.1



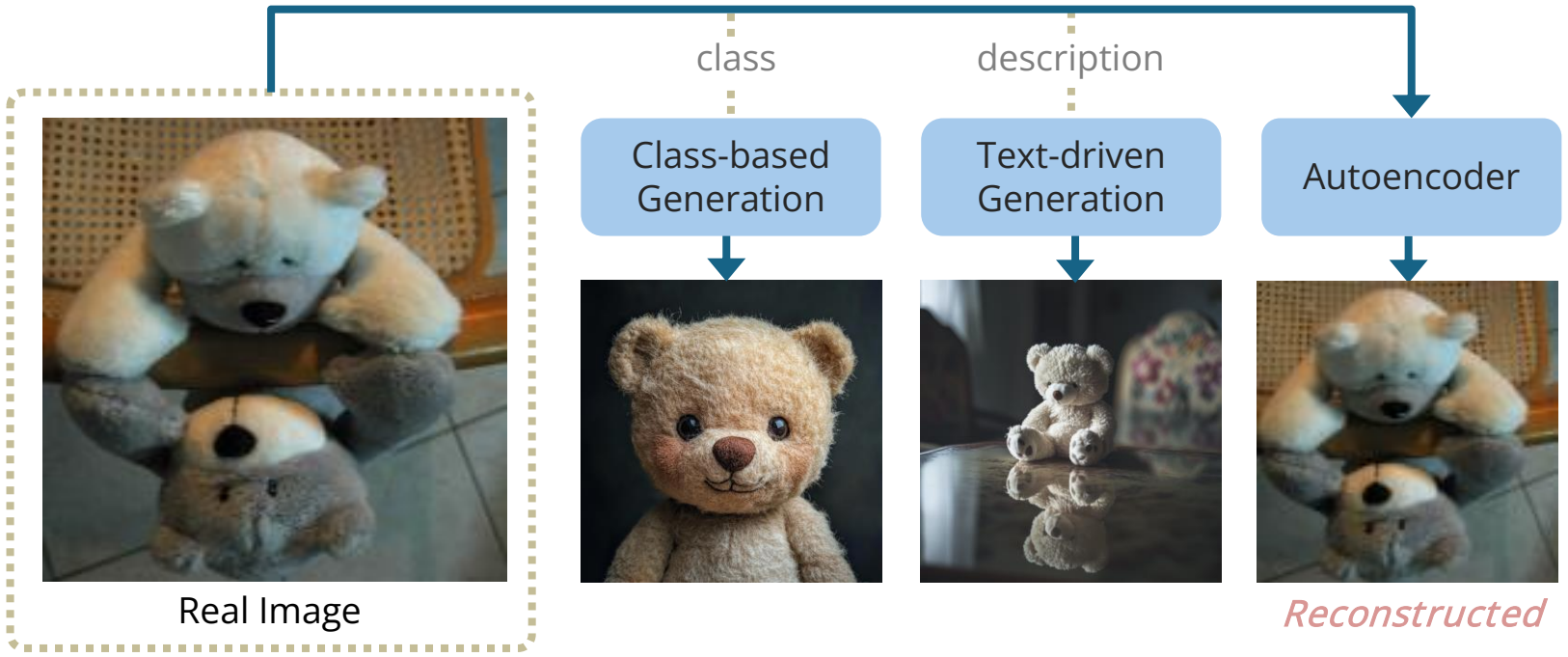
Proposed approach

- Our new training paradigm relies on the use of:
  - **self-conditioned reconstructions** of real images using Stable Diffusion 2.1
  - **content augmentation:** locally inpainted versions of the images



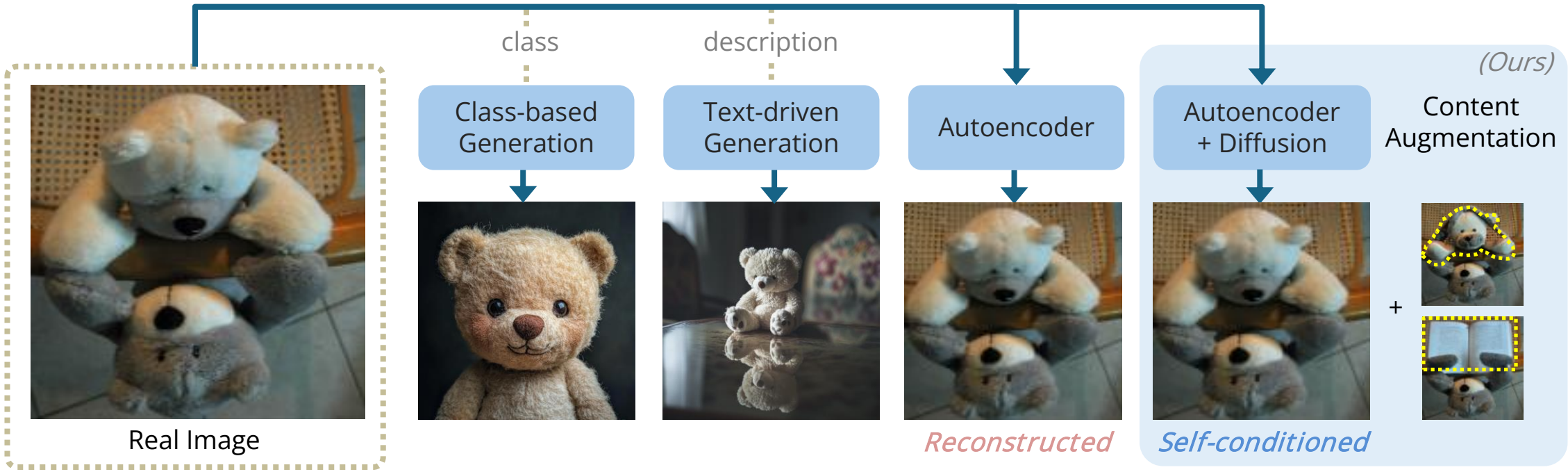
Self-conditioning vs previous approaches

- The **self-conditioned** image shares the same content as the real one
- **Diffusion steps** are also included to keep the forensic artifacts of the generation process



Self-conditioning vs previous approaches

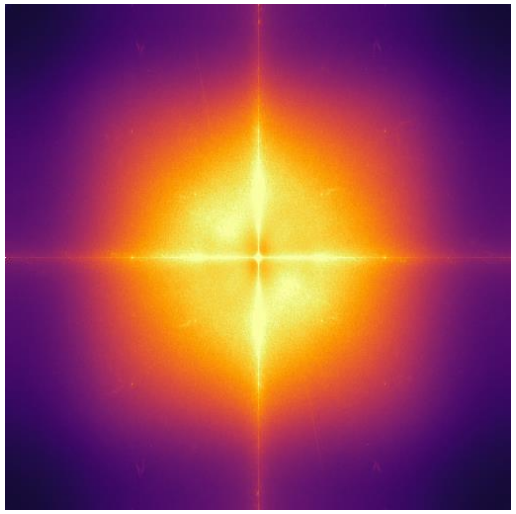
- The **self-conditioned** image shares the same content as the real one
- **Diffusion steps** are also included to keep the forensic artifacts of the generation process



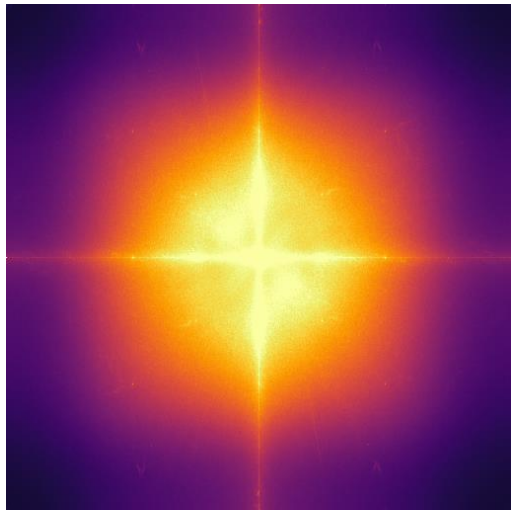
Frequency-based analysis

- The use of diffusion steps allows to exploit the inconsistencies across a wider range of frequencies
- Self-conditioned images embed forensics artifacts even at lower frequencies

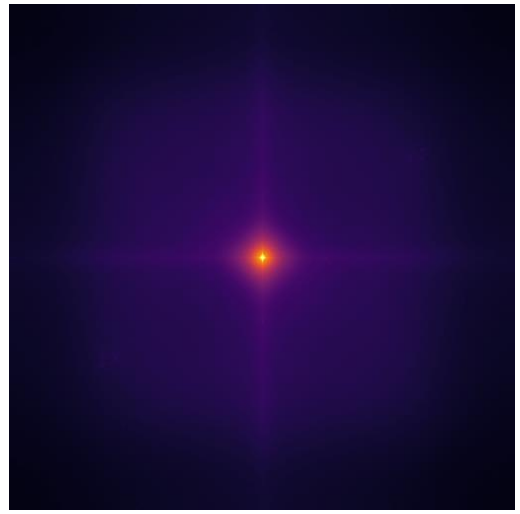
Real – Reconstructed



Real – Self-conditioned



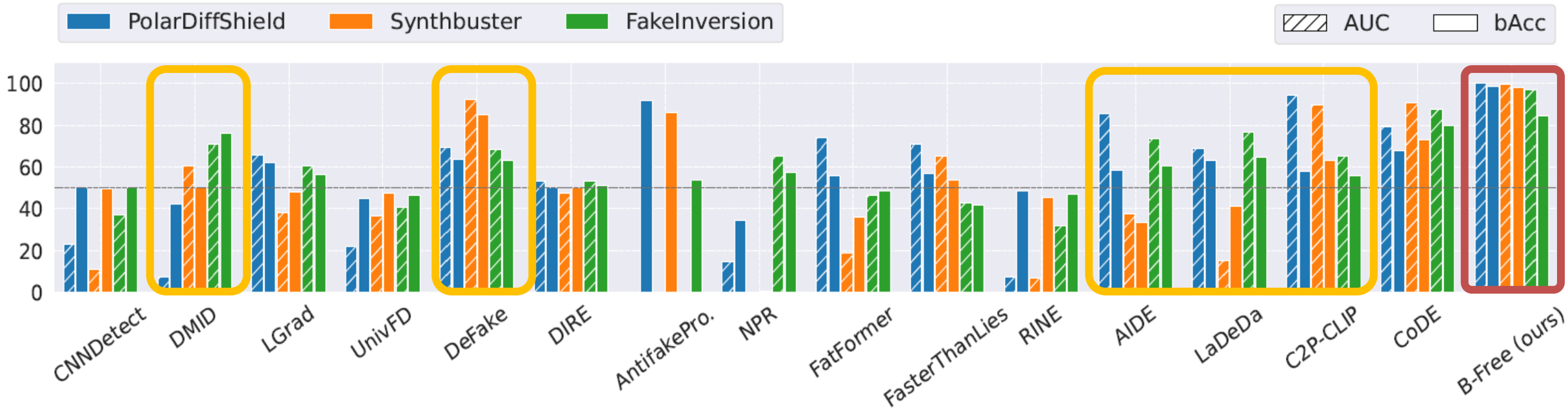
Reconstructed – Self-conditioned



Power spectra
(by averaging the **difference** on 2000 images)

Comparison with SoTA

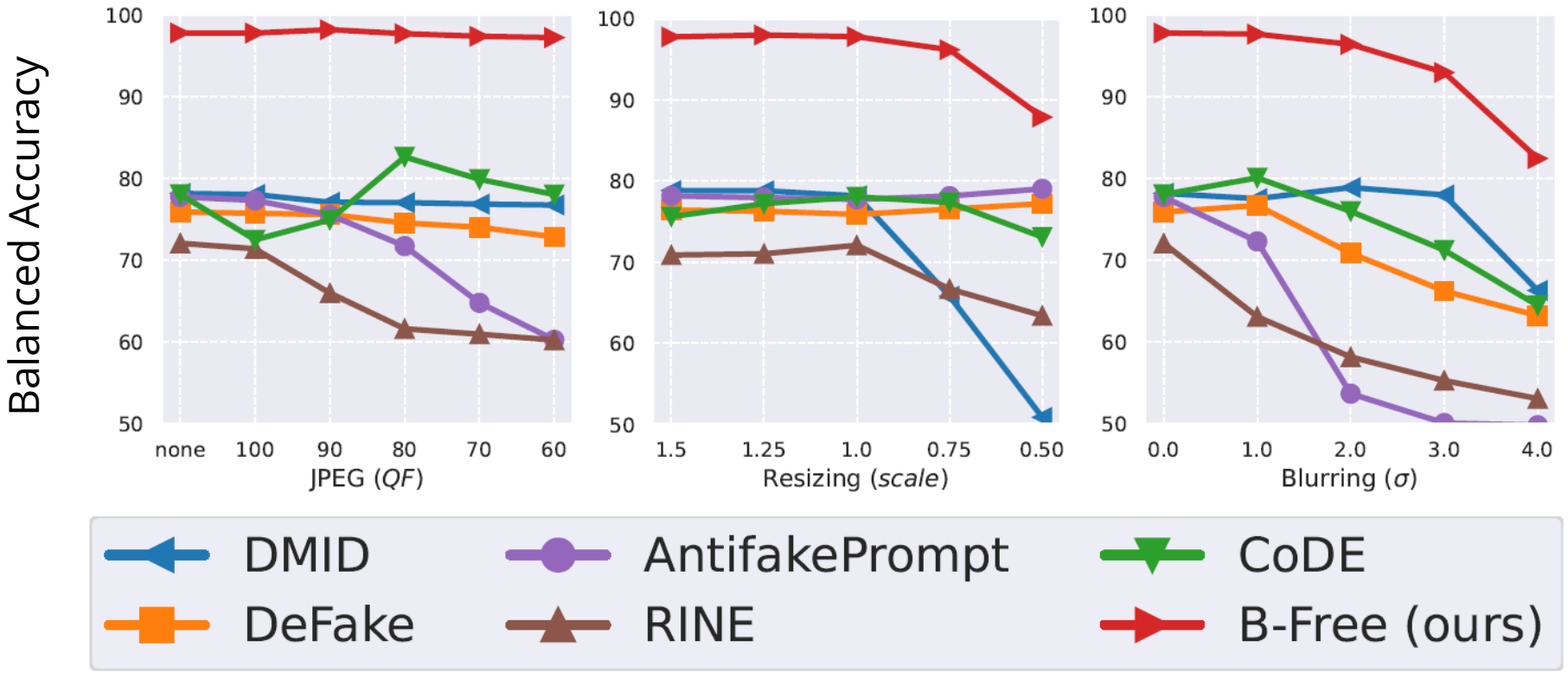
- Several methods do not have consistent performance on the same generator, varying even by 20% across datasets
- Our method provides much more stable predictions



Performance on DALL-E 3

Robustness analysis

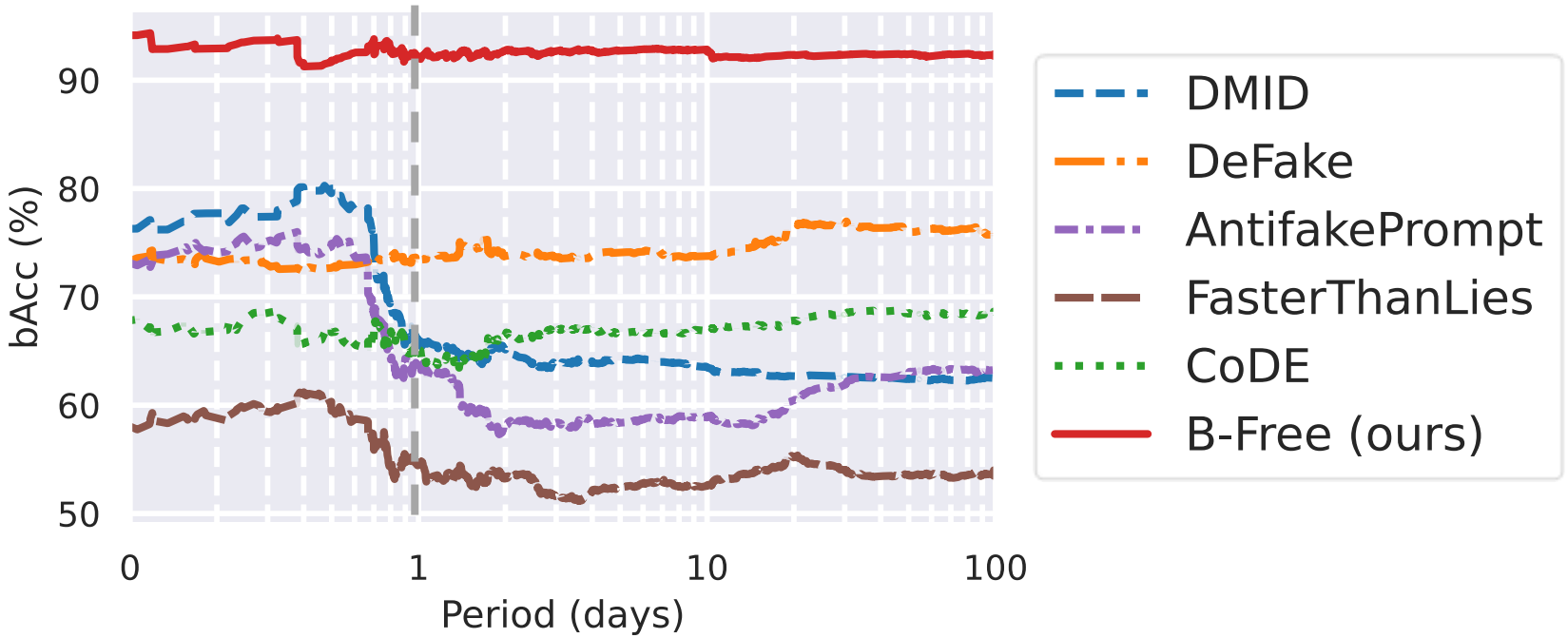
- We analyze SoTA performance under increasing levels of degradation, focusing on common operations that typically impact performance



Results on viral images

- We collected images that went viral and evaluated the performance in function of the time elapsed from the first online post
- Their quality is affected after just one day of re-posting
- **B-Free** performs always above 92%

Collection of **1400** real/fake images viral over the web



Conclusions

- We propose a **new training paradigm** for detecting AI-generated images and build a dataset of artificial fake images with the same semantic content of real images.

- We show that **content augmentation** further improves accuracy and calibration, particularly in real-world scenarios like social media

- Our findings suggest that thoughtful dataset design and training strategies can outperform current SoTA, encouraging **bias-free** approaches in forensic research

Thank you for your attention!