IMPACT OF AGGRESSIVE QUANTIZATION ON DEEPFAKE DETECTION

ANTONIO DI LAURO - 258788

CONTEXT AND MOTIVATION

- **The Trend:** Generative AI is moving to edge devices (smartphones, consumer GPUs).
- **The Enabler:** Quantization (FP16 -> FP8 -> FP4) drastically reduces VRAM usage and inference time.

The Research Question:

- Deepfake detectors rely on specific "fingerprints" (noise patterns, artifacts).
- Does aggressive quantization smooth out these fingerprints, acting as an unintentional adversarial attack?

EXPERIMENTAL SETUP (MODELS)

Generative Models (The Evolution):

- **Legacy:** Stable Diffusion 1.5, Stable Diffusion 2.1
- **High-Resolution**: SDXL
- **State-of-the-Art (Flow/DiT):** Stable Diffusion 3, Stable Diffusion 3.5, Flux.1

Quantization Levels:

- FP32 (Baseline), FP16, FP8, FP4

Dataset:

- 10,500 images generated (500 per configuration).

EXPERIMENTAL SETUP (DETECTORS)

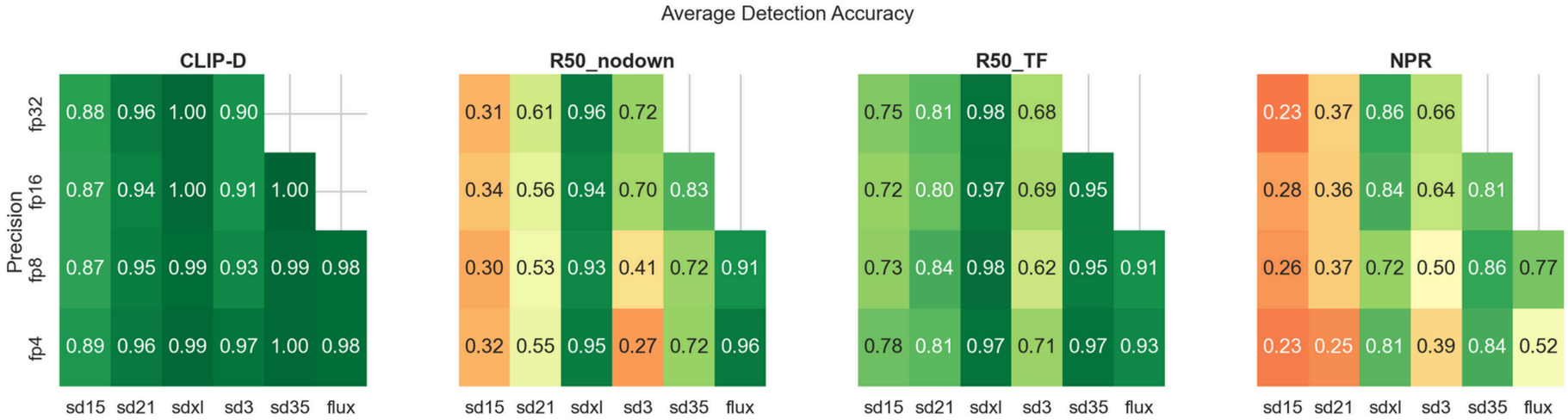
End-to-End (Trainable Backbone):

- Detectors: **NPR, ResNet50_nodown**
- Focus: Learned low-level artifacts / noise patterns.

Feature-Based (Frozen Backbone):

- Detectors: **ResNet50_TF, CLIP-D**
- Focus: High-level semantic anomalies and structural inconsistencies.

RESULTS OVERVIEW (HEATMAP)



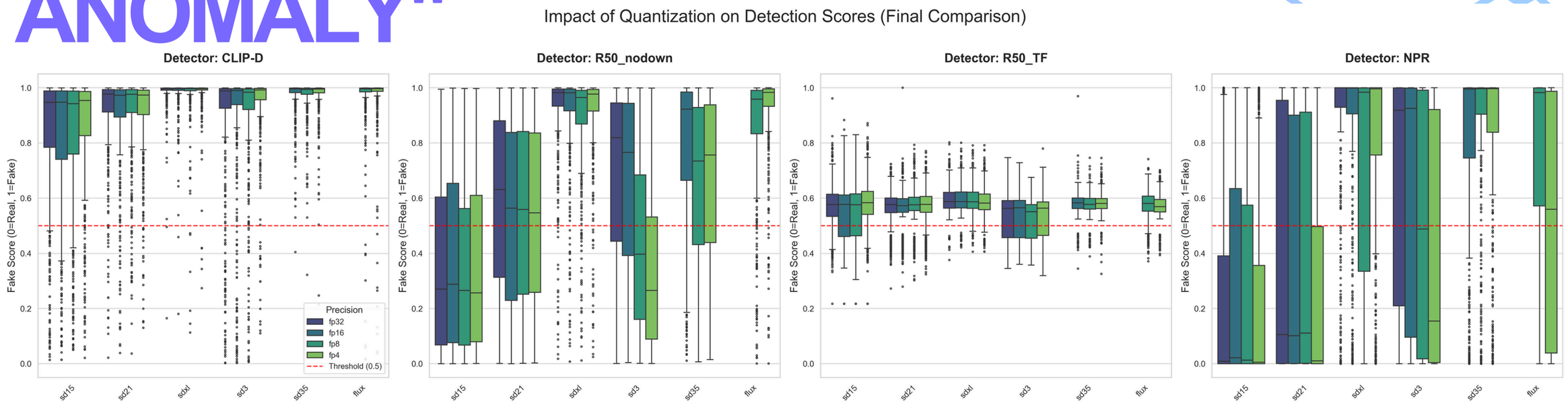
Green Zones:

SDXL and Flux are highly detectable across all precisions

Red Zones (The Failures):

- **Legacy Issues:** SD1.5 is challenging for modern detectors (NPR).
- **Quantization Issues:** SD3 at FP4 shows a significant drop in detection accuracy for End-to-End models.

THE "SD3 ANOMALY"



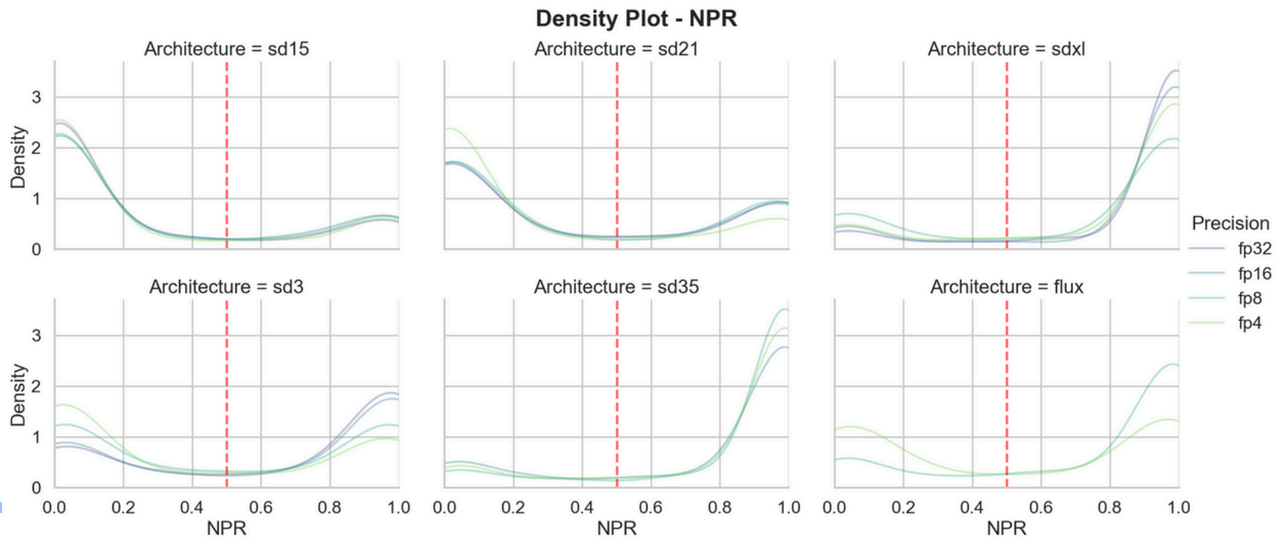
Stable Diffusion 3 (FP32 vs FP4):

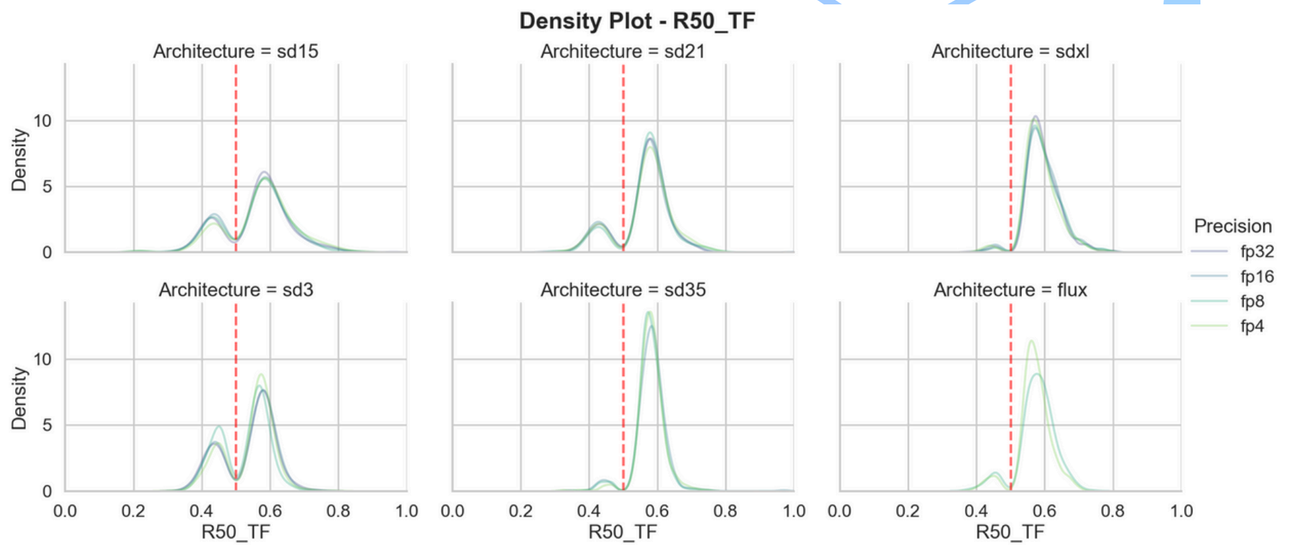
- R50_nodown / NPR: Accuracy drops significantly (from ~0.9 to <0.4).
- R50_TF: Accuracy remains stable (>0.9).

Interpretation:

Quantization acts as a "smoothing filter," removing the specific high-frequency artifacts that End-to-End detectors rely on.

SCIENTIFIC EVIDENCE (DISTRIBUTION SHIFT)





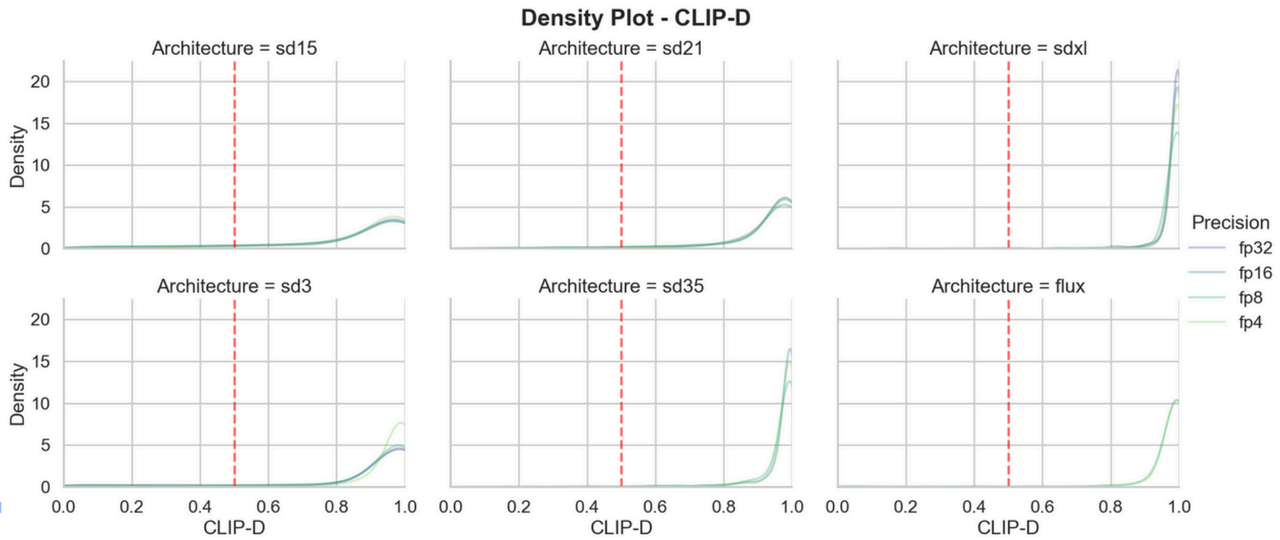
Deep Dive on NPR (SD3 Case):

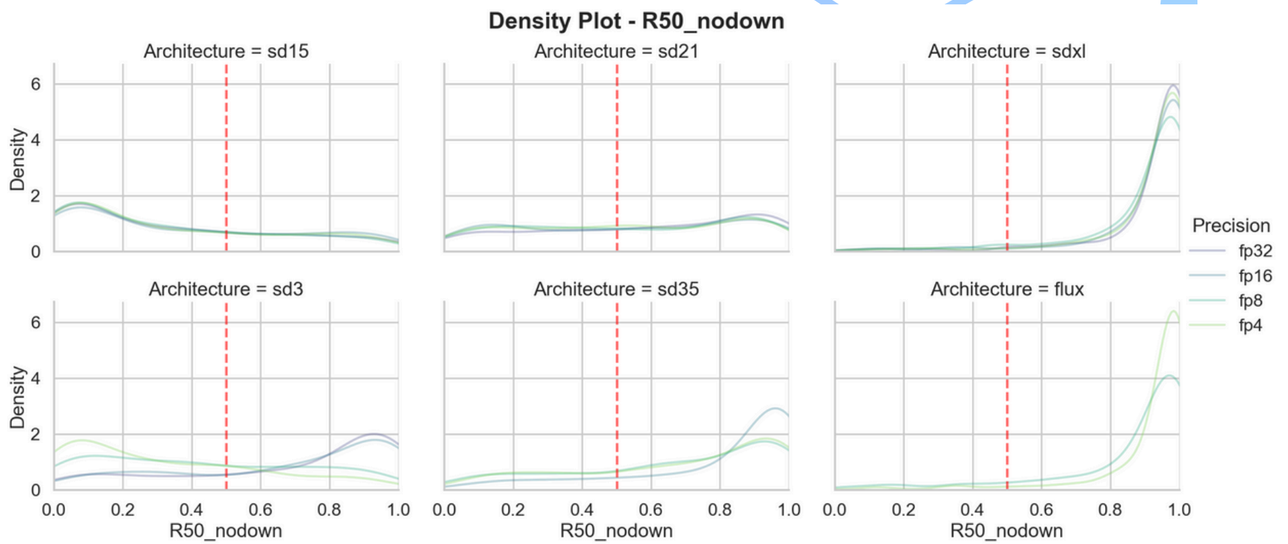
- **Purple Curve (FP32):** Peaked at 1.0 -> Detector is confident (True Positive).
- **Yellow/Green Curve (FP4/FP8):** Shifted to left/center -> Detector is confused (False Negative).

Comparison:

In contrast, R50_TF (Frozen) shows overlapping curves, indicating robustness.

ROBUSTNESS OF SOTA MODELS (FLUX & SDXL)





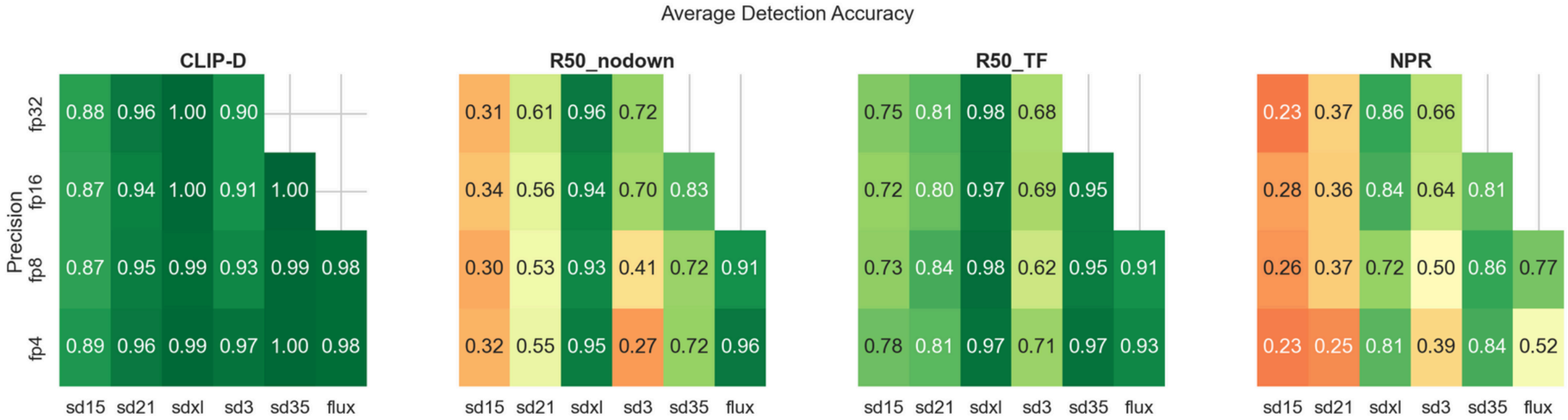
Observation:

Flux.1 and SDXL remain detectable (>95% accuracy) even at FP4.

Reasoning:

These models produce distinct structural fingerprints that are "stronger" than the noise reduction introduced by quantization.

THE LEGACY PROBLEM (SD 1.5)



Finding:

Modern detectors (NPR) perform poorly on older, lower-resolution generators (SD1.5).

Takeaway:

Detectors trained on recent high-res data suffer from **backward compatibility issues.**

CONCLUSION

Quantization as Obfuscation: FP4 quantization effectively removes specific artifacts, bypassing End-to-End detectors on architectures like SD3.

Detector Fragility: Detectors learning from scratch (NPR, R50_nodown) are less robust to compression than Feature-Based ones (R50_TF, CLIP).

Security Risk: As FP4 becomes standard for mobile inference, current "state-of-the-art" detectors may become obsolete without retraining.

THANK YOU

ANTONIO DI LAURO - 258788