SPECIOUS

SPECTRAL PERTURBATION ENGINE FOR CONTRASTIVE INFERENCE OVER UNIVERSAL SURROGATES

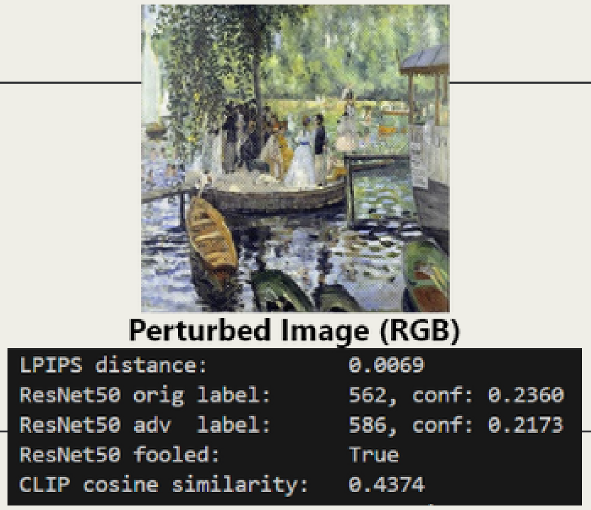
Dhruv Kumar, 00519011622 (LE)
Harshveer Singh, 07519011621
Deepanshu Singh, 01419011621

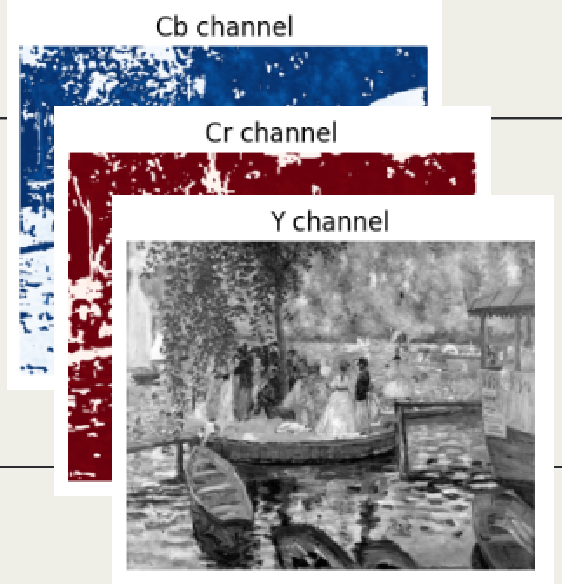
A universal, multi-model defensive engine that embeds imperceptible **high-frequency** perturbations into the **luminance (Y)** channel of YCbCr color space, which remain **invisible** to humans but degrade **image features** across multiple surrogate models (ResNet-50, CLIP ViT-B/32, etc.).

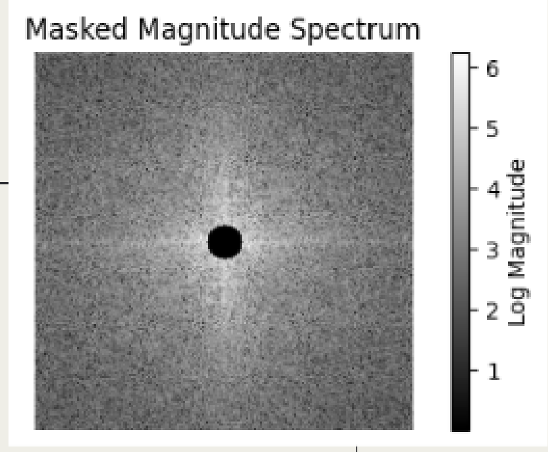
BACKGROUND AND MOTIVATION

- "**Ghiblification**": users turned photos into Studio Ghibli–style art overnight, igniting copyright debates.
- **Models train on vast, uncurated image sets**—often containing copyrighted works—without artists' consent.
- Copyright law covers specific images but generally not an **artist's "style**," leaving visual style unprotected.

AT A GLANCE







$$L_{SPECIOUS} = \exp\left(score_{lpips}(x, \hat{x}) - \left(\beta_{resnet} \left|\left| \Phi_{resnet}(x) - \Phi_{resnet}(\hat{x}) \right|\right|^2 + \beta_{clip} \left|\left| \Phi_{clip}(x) - \Phi_{clip}(\hat{x}) \right|\right|^2 \right) \right) + \lambda \max\left(0,\ score_{lpips}(x, \hat{x}) - \tau\right)$$

Universal Adversarial Perturbation

label- and model-agnostic, works across state-of-the-art architectures

YCbCr Color Space

Perturbations are added in the Luminosity (Y) channel of YCbCr color space, as adversarial perturbations prevail in it.

High Frequency in Fourier Domain

Targeted high frequencies in the Fourier domain, as these are the sharp edges and textures that AI models use to learn.

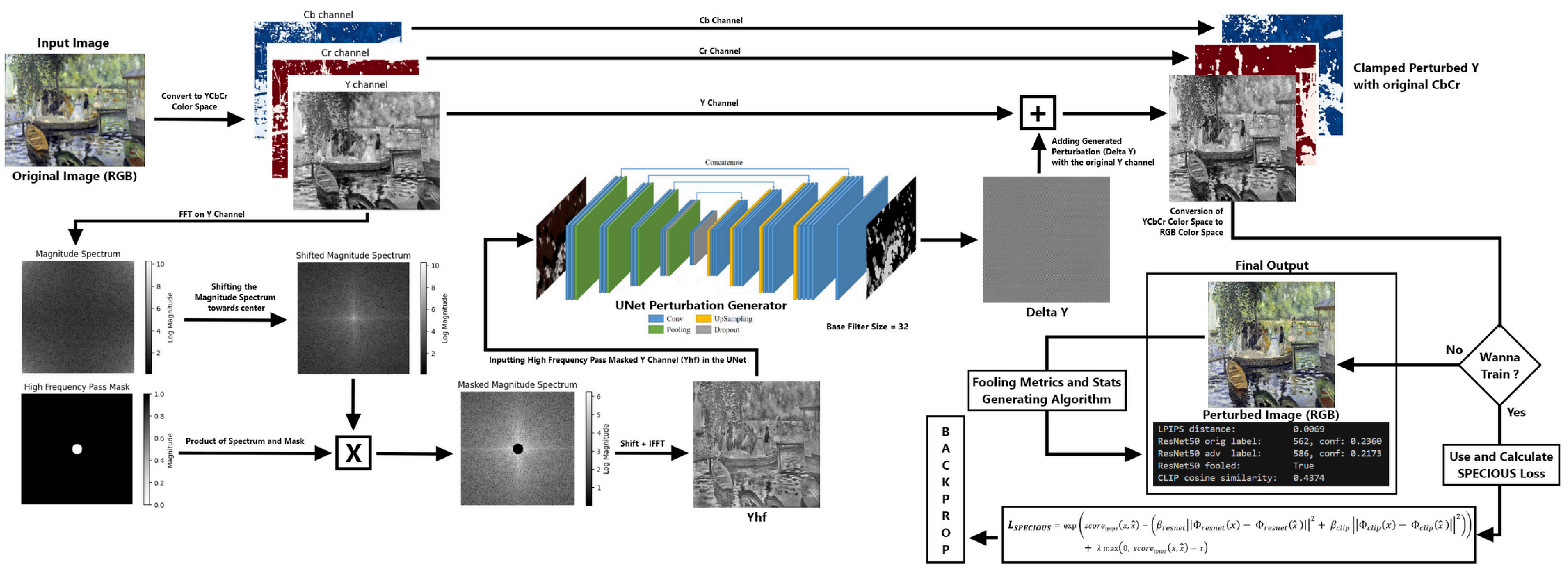
SPECIOUS Loss Function

Joint loss function which minimizes LPIPS (perceptual similarity score) + maximizes feature distortion across surrogate models.

SPECIOUS

METHODOLOGY

SPECIOUS (Spectral Perturbation Engine for Contrastive Inference Over Universal Surrogates)



DESIGNING OF THE NOVEL SPECIOUS LOSS

Perceptual Similarity (LPIPS):

$$\boldsymbol{score_{lpips}(x,\hat{x})} = lpips(x,\hat{x})$$

Feature Distortion (ResNet-50 & CLIP):

$$\boldsymbol{d_{feature}} = \beta_{resnet}||\Phi_{resnet}(x) - \Phi_{resnet}(\hat{x})||^2 + \beta_{clip}||\Phi_{clip}(x) - \Phi_{clip}(\hat{x})||^2$$

Strict Positivity:

$$\boldsymbol{L_{exp}} = \exp(score_{lpips}(x,\hat{x}) - d_{feature})$$

Imperceptible Penalty:

$$\boldsymbol{Penalty} = \lambda\, max(0,\ score_{lpips}(x,\hat{x}) - \tau)$$

Total Loss:

$$\boldsymbol{L_{SPECIOUS}} = \boldsymbol{L_{exp}} + \boldsymbol{Penalty}$$

RESULTS AND FINDINGS

We evaluated SPECIOUS on both **classification (ResNet-50)** and **zero-shot retrieval (CLIP ViT-B/32)** tasks, as well as analyzed training dynamics. All experiments use our **10,000 image** corpus **(5k Pascal VOC + 5k Artworks)** at **224** × **224**, **base_filters**=**32**, trained for **7 epochs**, and for testing, we used the well-curated test set of **2,000 images** (1k Pascal VOC + 1k Artworks)

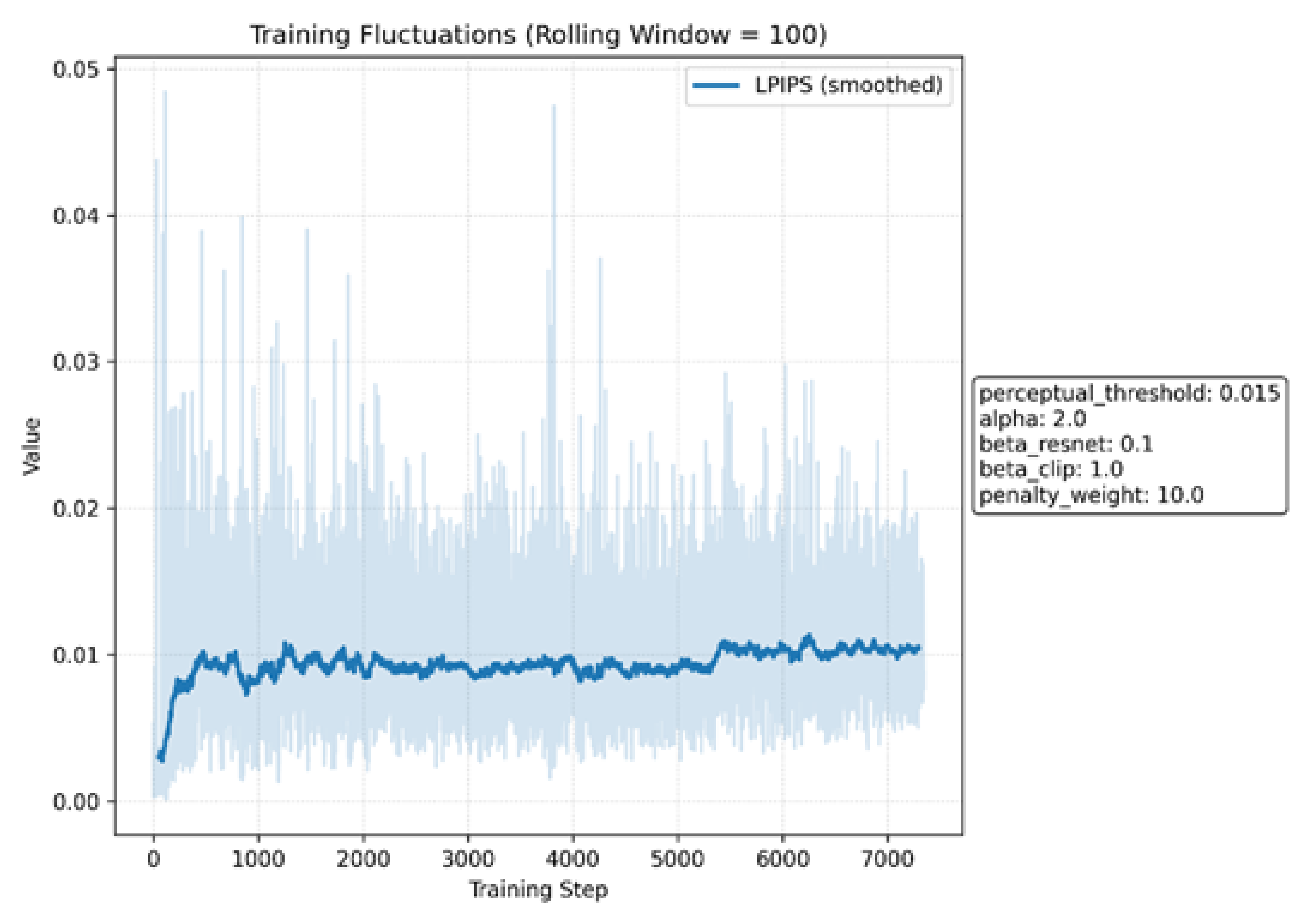
TRAINING DYNAMICS

- **Feature loss (orange)** increases sharply in the first 1,000 steps, driven largely by CLIP embedding distortion (β_clip = 5.0), and plateaus near **0.20–0.22**.
- **Total loss (green)** smoothly decreases, settling at **~ 0.83** by the end of training, indicating a balance between perceptual and feature objectives.

SPECIOUS

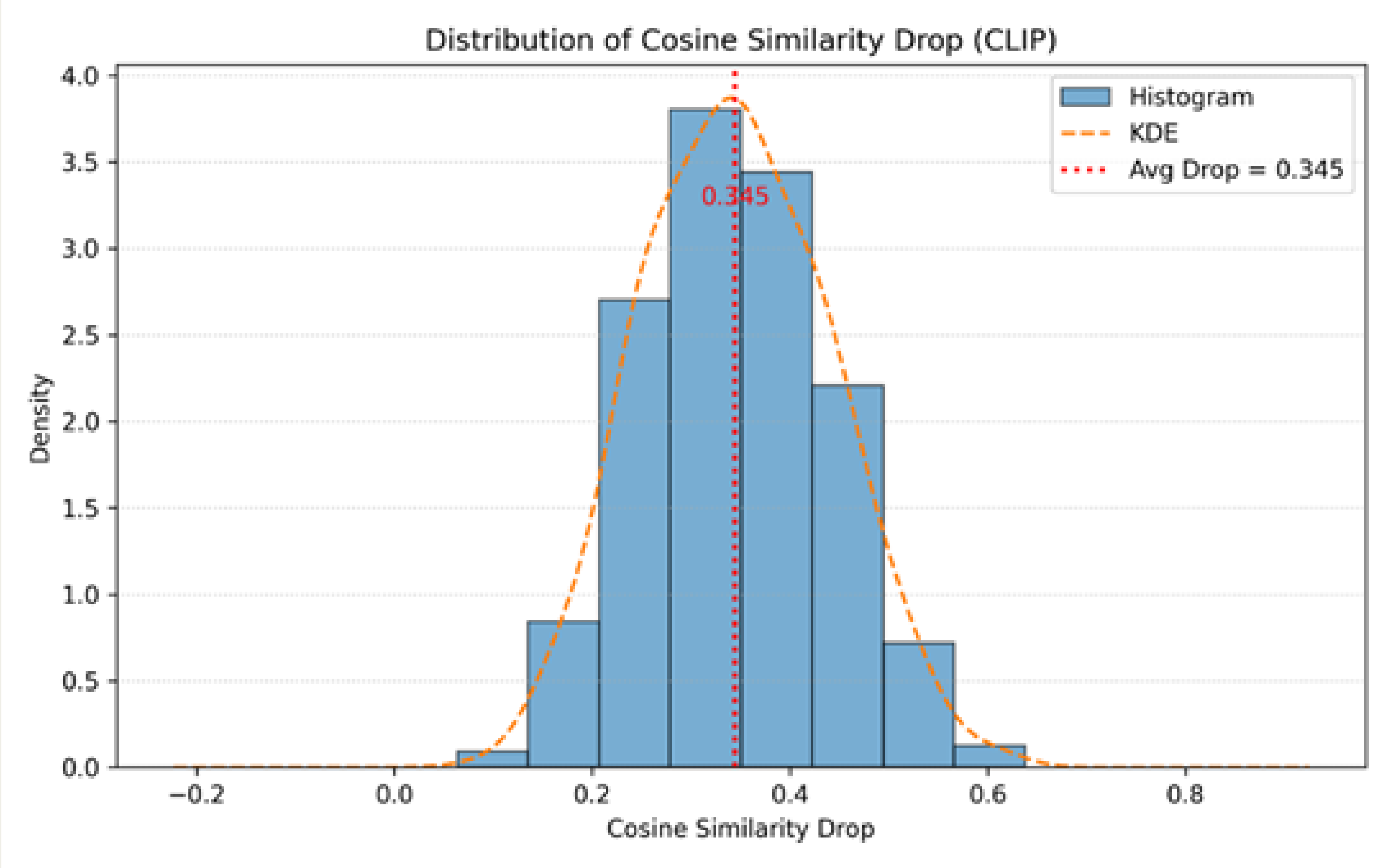
TRAINING DYNAMICS II

LPIPS (blue) quickly rises from near zero to **~0.01** within 500 steps, then stabilizes around **0.009–0.011**, well below our threshold **τ = 0.015**

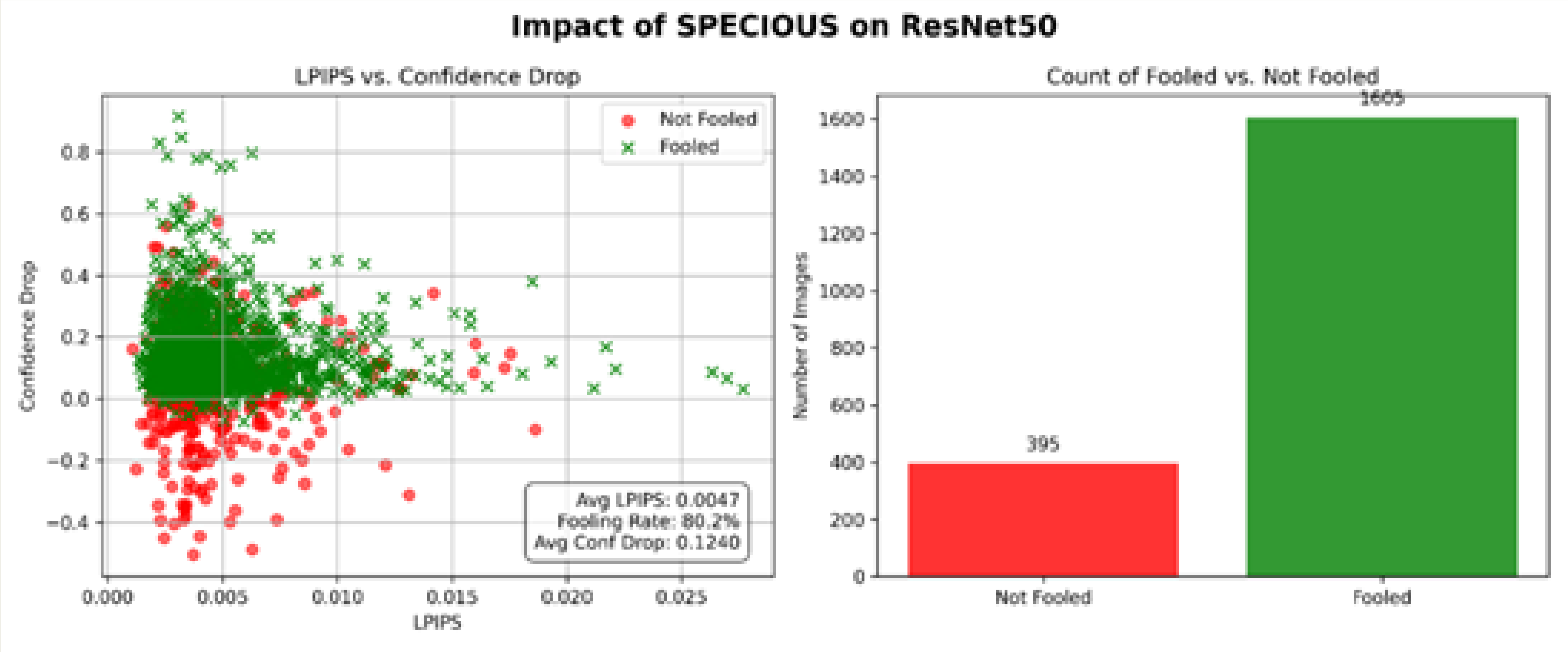


CLIP COSINE-SIMILARITY DROP DISTRIBUTION

- Average drop of **0.345** in cosine similarity indicates substantial embedding shift.
- The distribution is roughly **Gaussian** with a very **less standard deviation**, with most drops between **0.25–0.45**, confirming consistent disruption across samples.

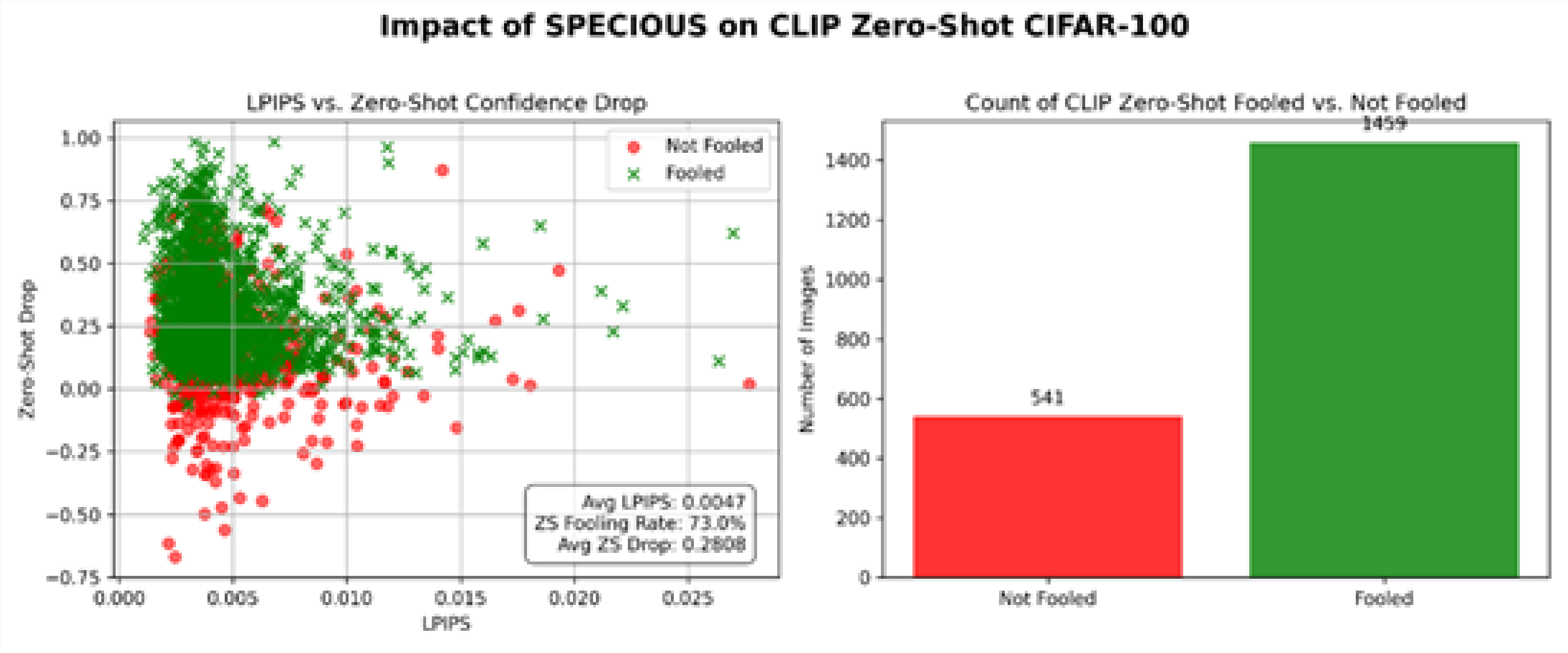


RESNET-50 CLASSIFICATION RESULTS



- **Fooling Rate: 80.2%** of images have top-1 labels flipped under ResNet-50.
- **Avg. LPIPS: 0.0047**, far below τ, demonstrating imperceptibility.
- **Avg. Confidence Drop: 0.1240**, indicating a meaningful reduction in model certainty.

CLIP ZERO-SHOT PREDICTION



- **Zero-Shot Fooling Rate: 73.0%** of images change their top-1 zero-shot label post-perturbation.
- **Avg. ZS Confidence Drop: 0.2808**
- **Avg. LPIPS: again 0.0047**, confirming consistency across tasks.

COMPETITIVE ANALYSIS

Limitation	Glaze	Nightshade	SPECIOUS
Model-Specific	diffusion-only	prompt-specific	multi-model
Target-Label Dependency	artist-preset style only	exact prompt/class only	label-agnostic
No LPIPS Minimization	no direct LPIPS control	no direct LPIPS control	bi-objective (LPIPS+feat-dist)
RGB-Only	spreads RGB channels	spreads RGB channels	Y-channel only
No Frequency Filter	spatial-only	spatial-only	learnable high-pass mask

CONCLUSION

In this work, we introduced **SPECIOUS ("Spectral Perturbation Engine for Contrastive Inference Over Universal Surrogates")**, a novel defence mechanism that injects **imperceptible, high-frequency perturbations** into the **Y channel** of images to disrupt multiple black-box encoders simultaneously. By combining a **learnable high-pass mask** in the **Fourier domain** with a **U-Net generator**, SPECIOUS focuses its perturbations on **edges and textures**—features to which deep models are most sensitive. Training with our **Specious Loss**, which **minimizes LPIPS** (perceptual similarity) while **maximizing squared-error feature distortion** on pre-trained ResNet-50 and CLIP ViT-B/32 embeddings, yields perturbations that are **nearly invisible to humans** (**LPIPS < 0.01**) yet cause significant embedding shifts (**avg. CLIP cosine drop = 0.345**) and **high fooling rates** (**>80% on ResNet-50, >70% on CLIP zero-shot**).

Thank you!

Dhruv Kumar, 00519011622 (LE)
Harshveer Singh, 07519011621
Deepanshu Singh, 01419011621

SPECIOUS