UMassAmherst

Manning College of Information & Computer Sciences

IBM

Georgia Tech

NEURAL INFORMATION PROCESSING SYSTEMS

Attack-Resilient Image Watermarking Using Stable Diffusion

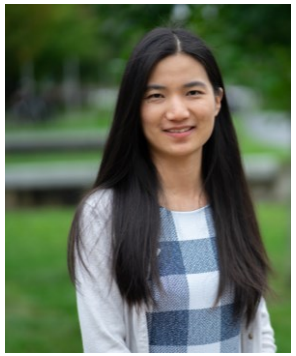
Lijun Zhang
UMass

Xiao Liu
UMass

Antoni Viros Martin
IBM

Cindy Xiong Bearfield
Georgia Tech

Yuriy Brun
UMass



Hui Guan
UMass

Common Image Watermarking

Develop image watermarking method to inject **invisible and robust** watermark into the given image

Image Watermarking

Existing Methods

(e.g., traditional frequency decomposition, NN-based methods)

Trade-off: Image quality ⇔ Watermark robustness

Method	Image Quality SSIM	Post-Attack (JPEG) WDR
SSL	**0.98**	0.046
StegaStamp	0.91	**1.0**

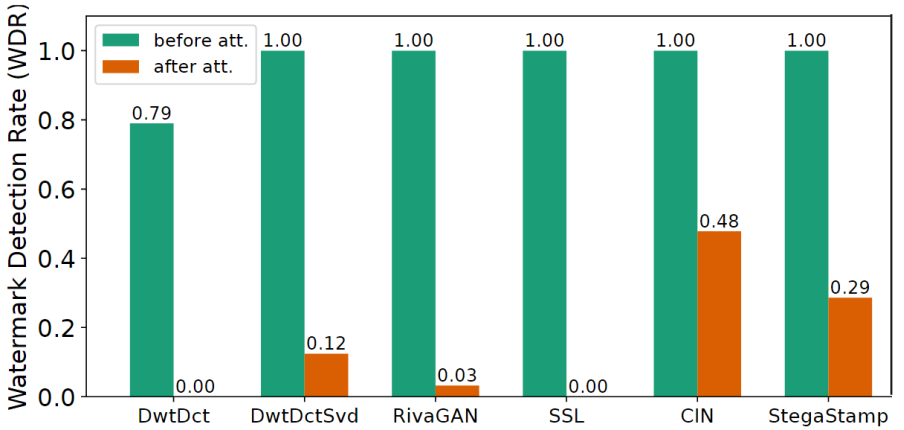
Invisibility

Image quality after watermarking (e.g., SSIM, PSNR)

Robustness

Watermark detection rate (i.e., WDR) after being attacked

New Challenge: Stable diffusion-based watermark removal[1]



[1] Xuandong Zhao, etc. Generative autoencoders as watermark attackers: Analyses of vulnerabilities and threats. ICML Workshop 2023.

Image Watermarking

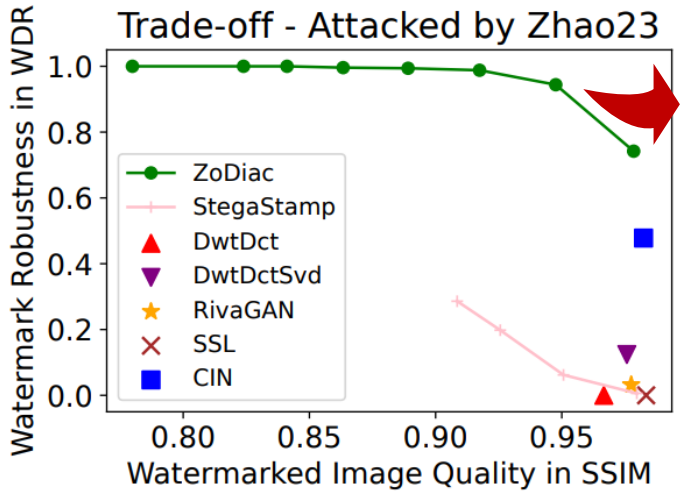
Invisibility

Image quality after watermarking (e.g., SSIM, PSNR)

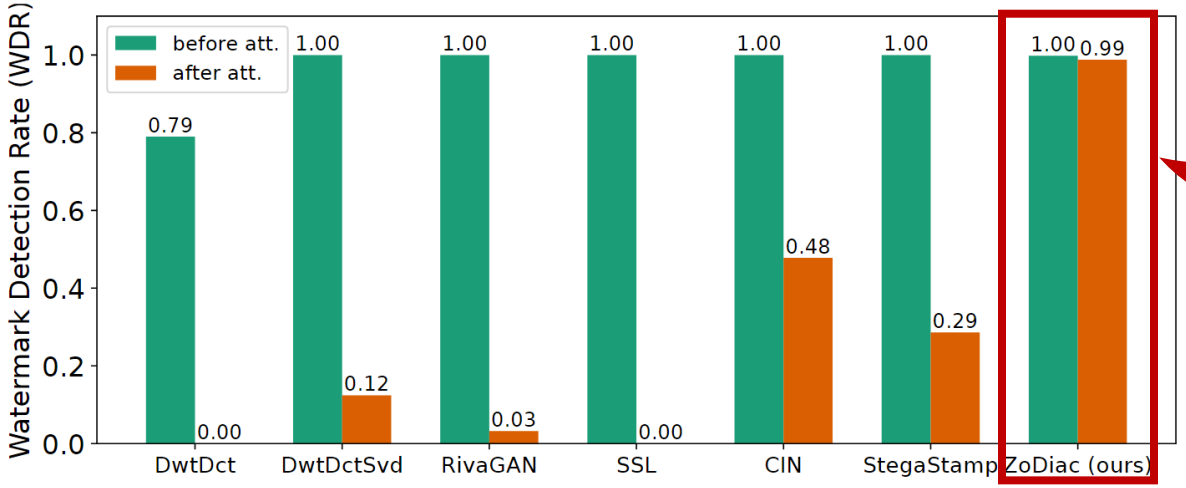
Robustness

Watermark detection rate (i.e., WDR) after being attacked

Proposed Method: ZoDiac

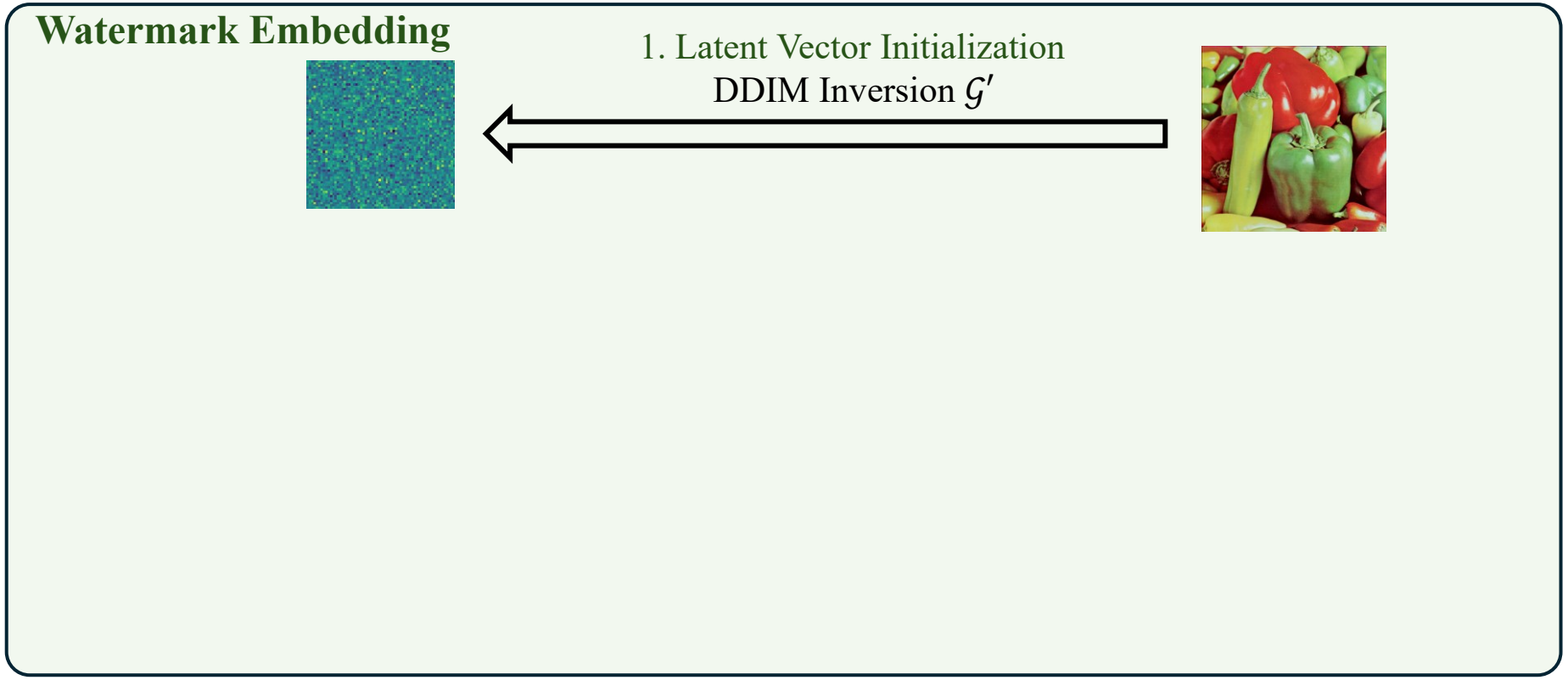


Adjustable and Robust trade-off



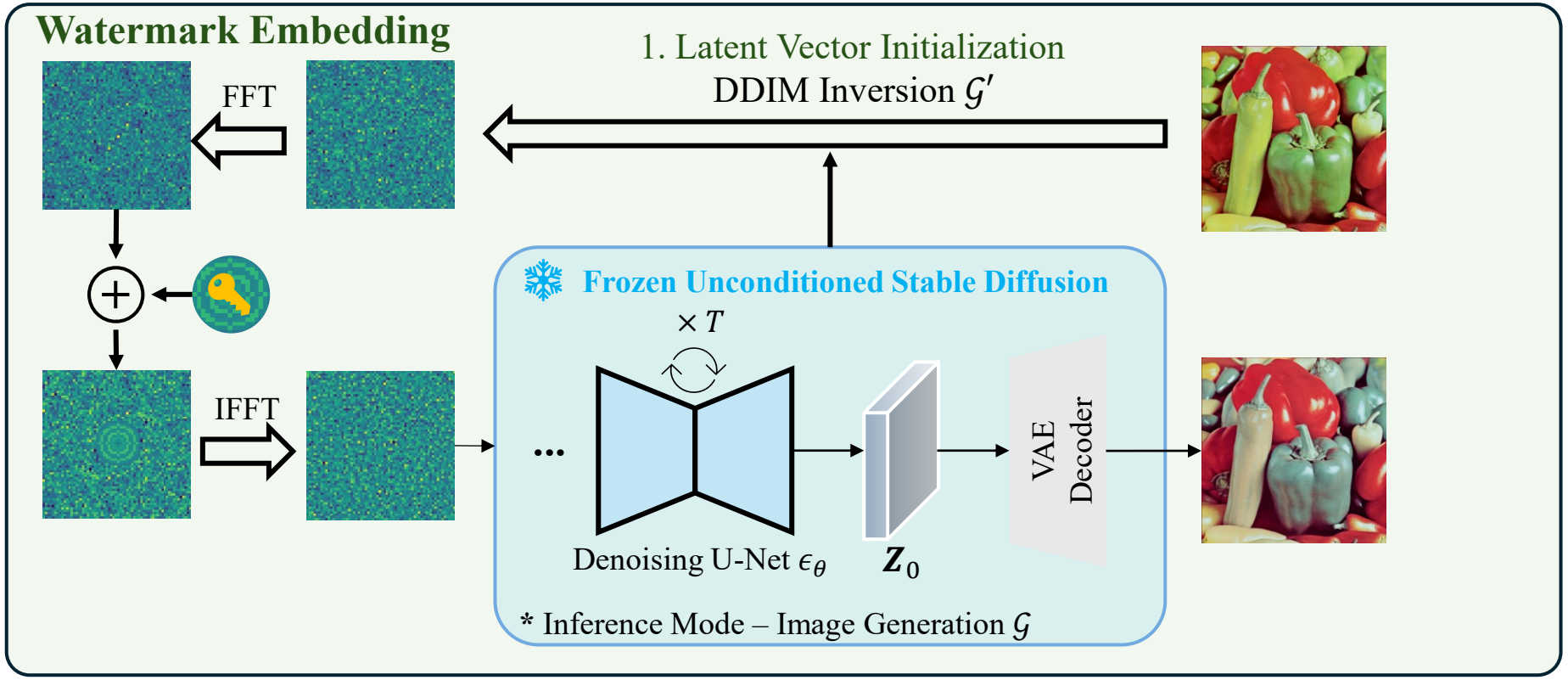
Robust to Strong Attack

ZoDiac Framework



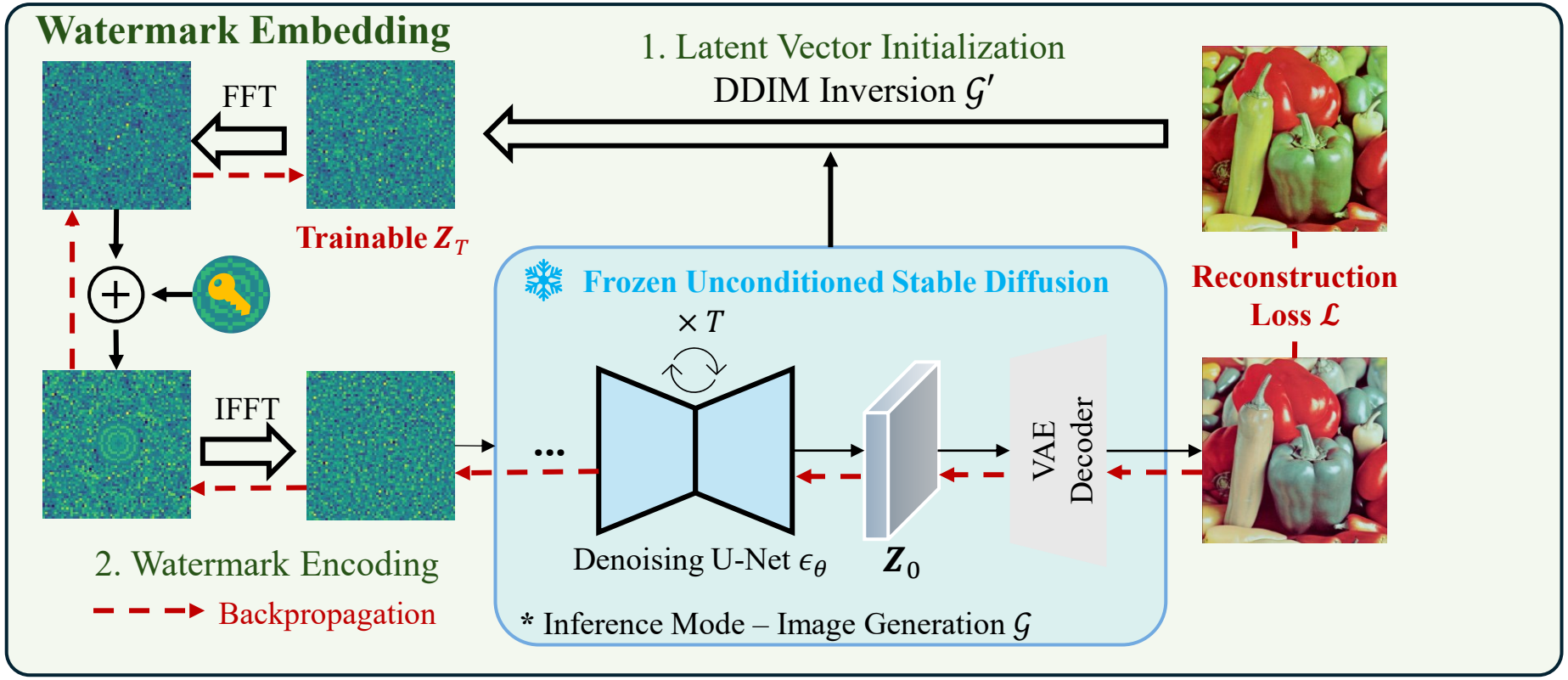
Learn *a latent vector that encodes a pre-defined watermark within its Fourier space, and can be mapped by pre-trained stable diffusion models into an image closely resembling the original image.*

ZoDiac Framework



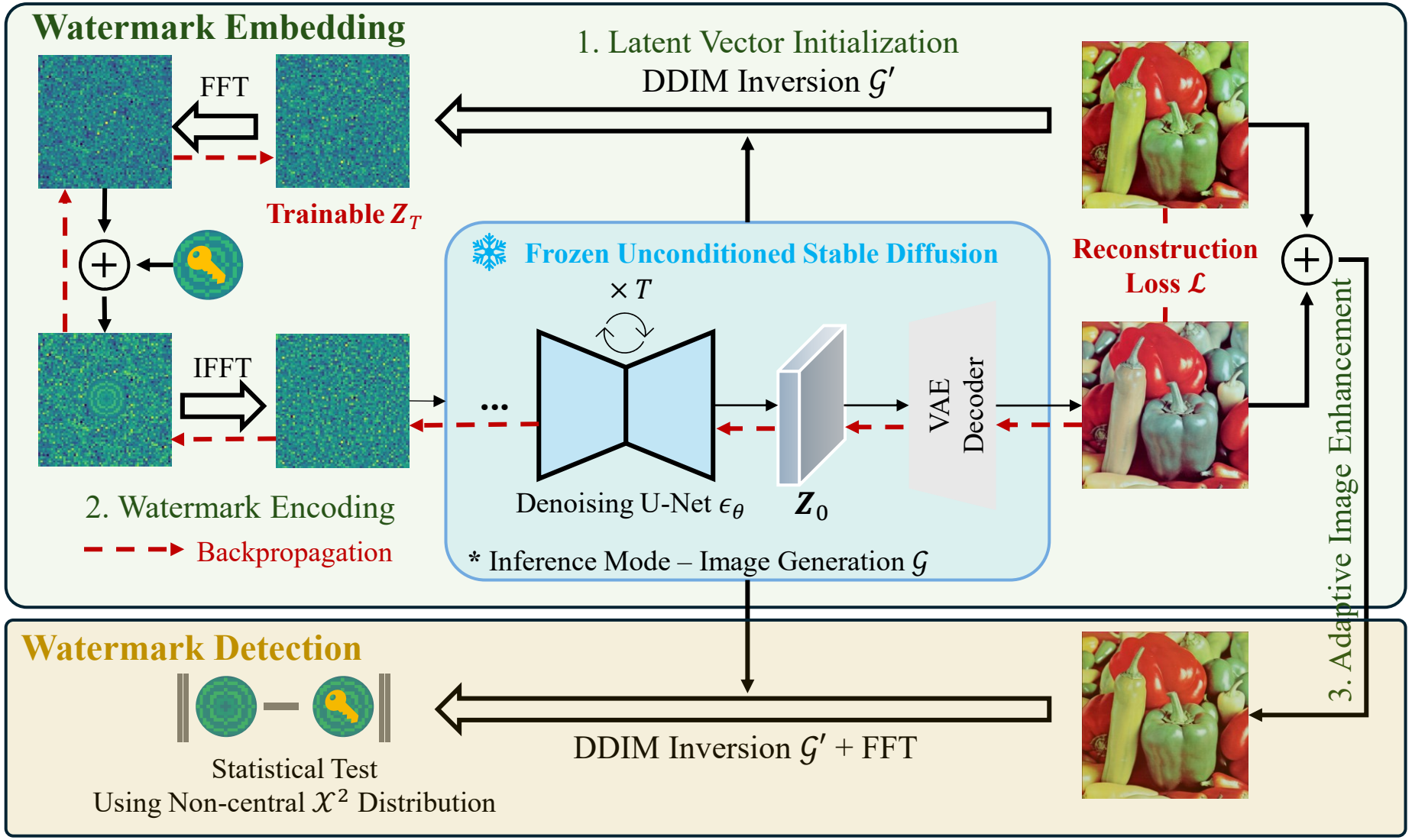
Learn *a latent vector that encodes a pre-defined watermark within its Fourier space, and can be mapped by pre-trained stable diffusion models into an image closely resembling the original image.*

ZoDiac Framework

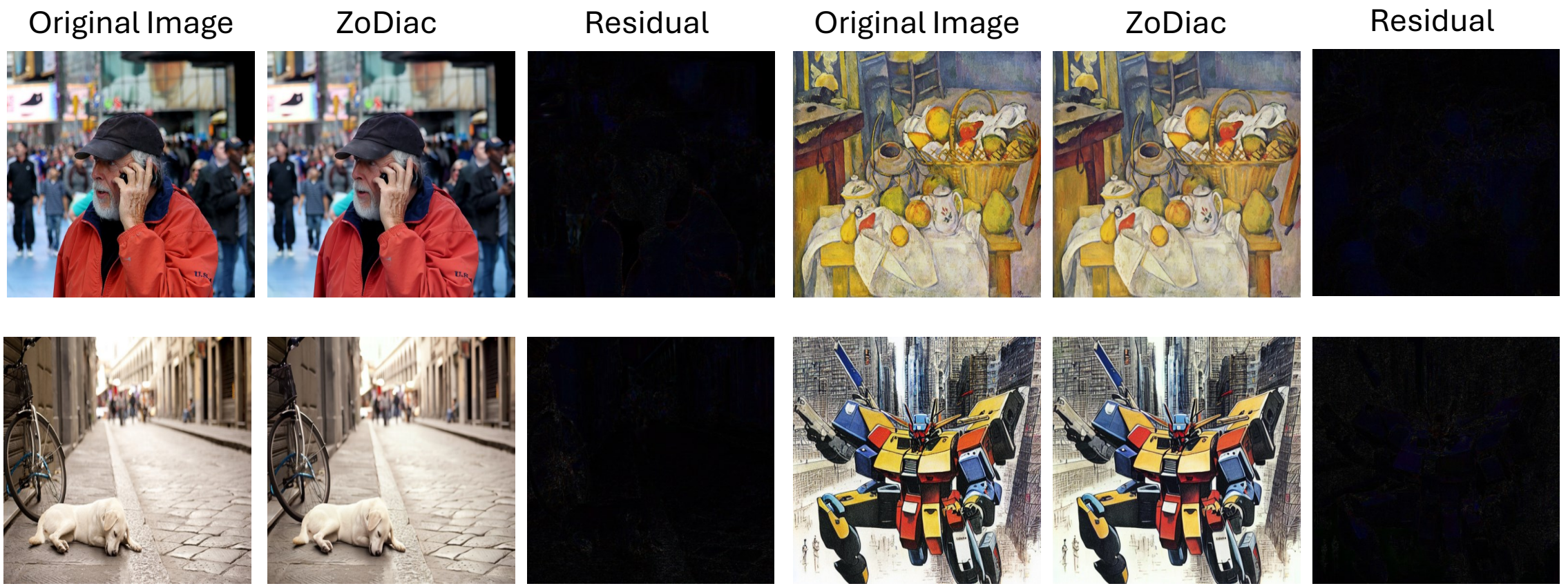


Learn *a latent vector that encodes a pre-defined watermark within its Fourier space, and can be mapped by pre-trained stable diffusion models into an image closely resembling the original image.*

ZoDiac Framework



Visual Examples



No significant visual influence after injecting watermark

Can handle image with different domains

Takeaway

- A novel framework for embedding invisible watermarks into existing images using *any* pre-trained stable diffusion.
- Strong robustness against most watermark attacks with watermark detection rate above 98% and false positive rate below 6.4%.
- Resilient to generative-AI based image regeneration with WDR above 98% while baselines fails below 50%.

UMassAmherst
Manning College of Information & Computer Sciences

IBM

Georgia Tech

NEURAL INFORMATION PROCESSING SYSTEMS

Attack-Resilient Image Watermarking Using Stable Diffusion

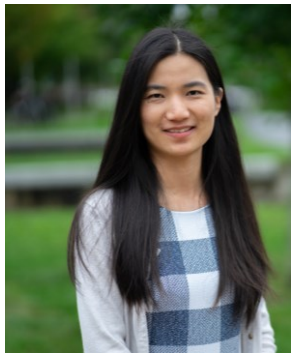
Lijun Zhang
UMass

Xiao Liu
UMass

Antoni Viros Martin
IBM

Cindy Xiong Bearfield
Georgia Tech

Yuriy Brun
UMass



Hui Guan
UMass