CVPR
JUNE 17-21, 2024
SEATTLE, WA

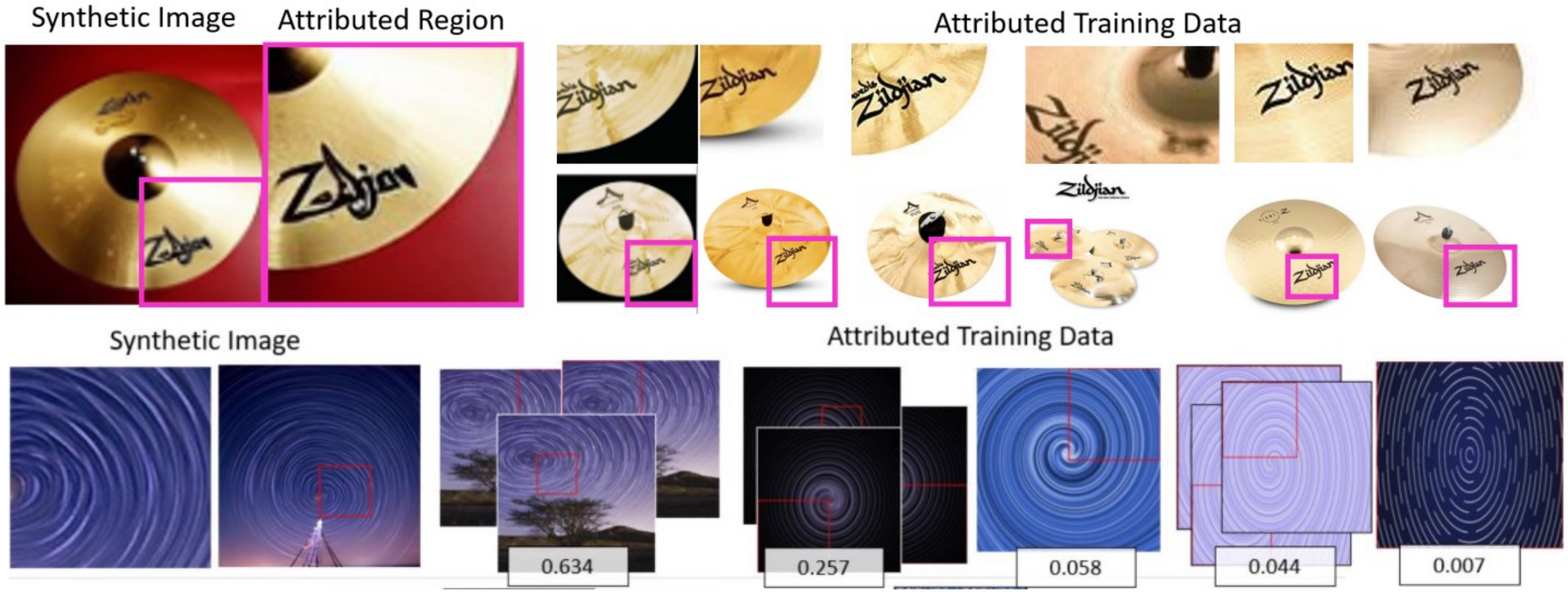
ProMark: Proactive Diffusion Watermarking for Causal Attribution

Vishal Asnani[1,2] , John Collomosse[2,3] , Tu Bui[3] , Xiaoming Liu[2] , Shruti Agarwal[1]
[1]Michigan State University, [2]Adobe Research, [3]University of Surrey

Poster: Thursday morning, #: 111

Motivation: Concept Attribution

Generative AI (GenAI) **remembers motifs and styles** from training data while generating synthetic images.[1]



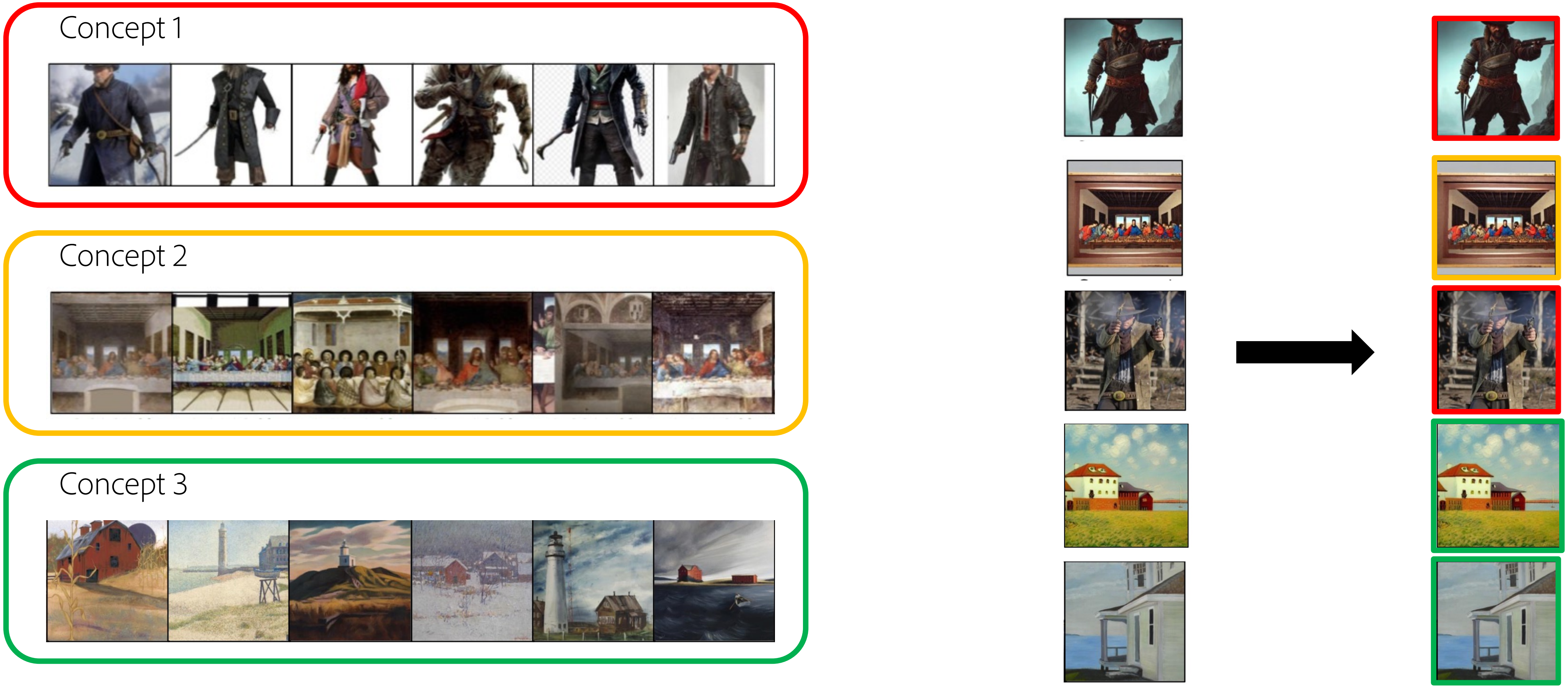
Concept attribution: Source training images should get credit if synthetic images are influenced by them.

[1] Balan, Kar, et al. "EKILA: Synthetic Media Provenance and Attribution for Generative Art." *In CVPRW* 2023.

Problem overview

Training data

Synthetic images



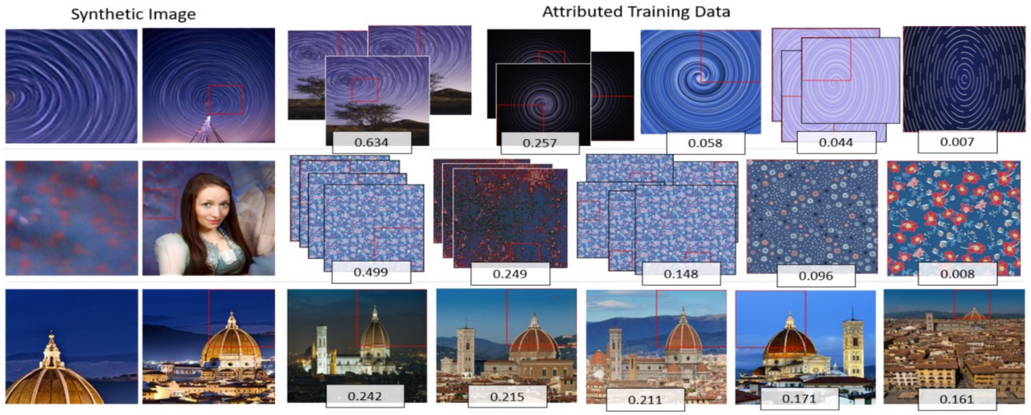
Concept 1

Concept 2

Concept 3

Passive ways to perform attribution

EKILA[1]

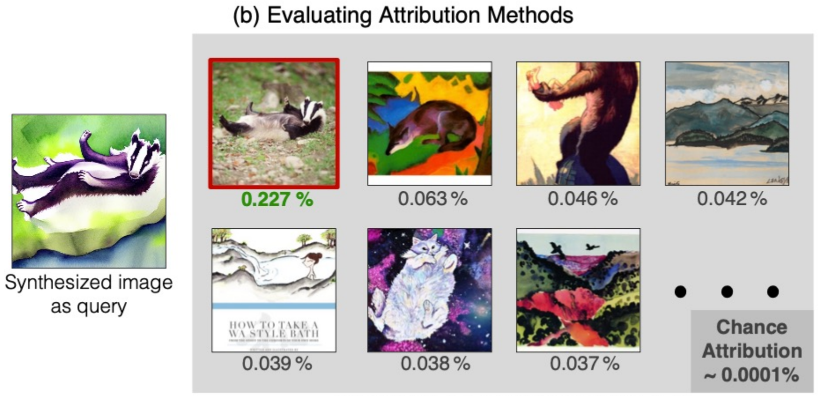


Matching the patch fingerprints to attribute them to training image patches

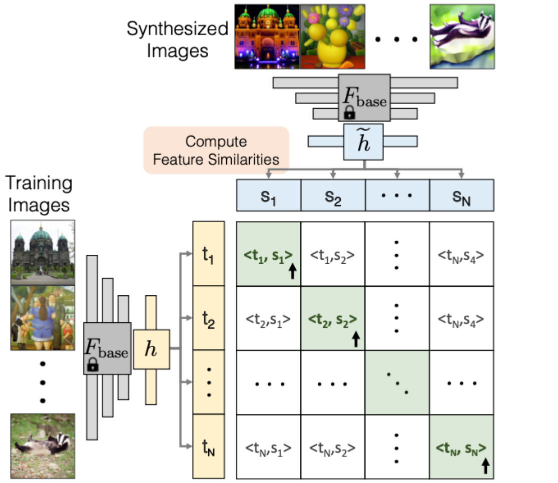
ALADIN[2]

Learns a representation for fine-grained style similarity

AbC[3]


(b) Evaluating Attribution Methods
Synthesized image as query
0.227 %
0.063 %
0.046 %
0.042 %
0.039 %
0.038 %
0.037 %
Chance Attribution ~ 0.0001%

Attribute the exemplar from training data


Synthesized Images
F_{base}
$\tilde{h}$
Compute Feature Similarities
Training Images
s_1 s_2 ... s_N
t_1 t_2 t_N
F_{base}
h
$<t_1, s_1>$ $<t_1,s_2>$ $<t_N,s_4>$
$<t_2,s_1>$ $<t_2, s_2>$ $<t_N,s_4>$
$<t_N,s_1>$ $<t_N,s_2>$ $<t_N, s_N>$

Contrastive Learning Across Two Views

Using CLIP embeddings to perform attribution to source images

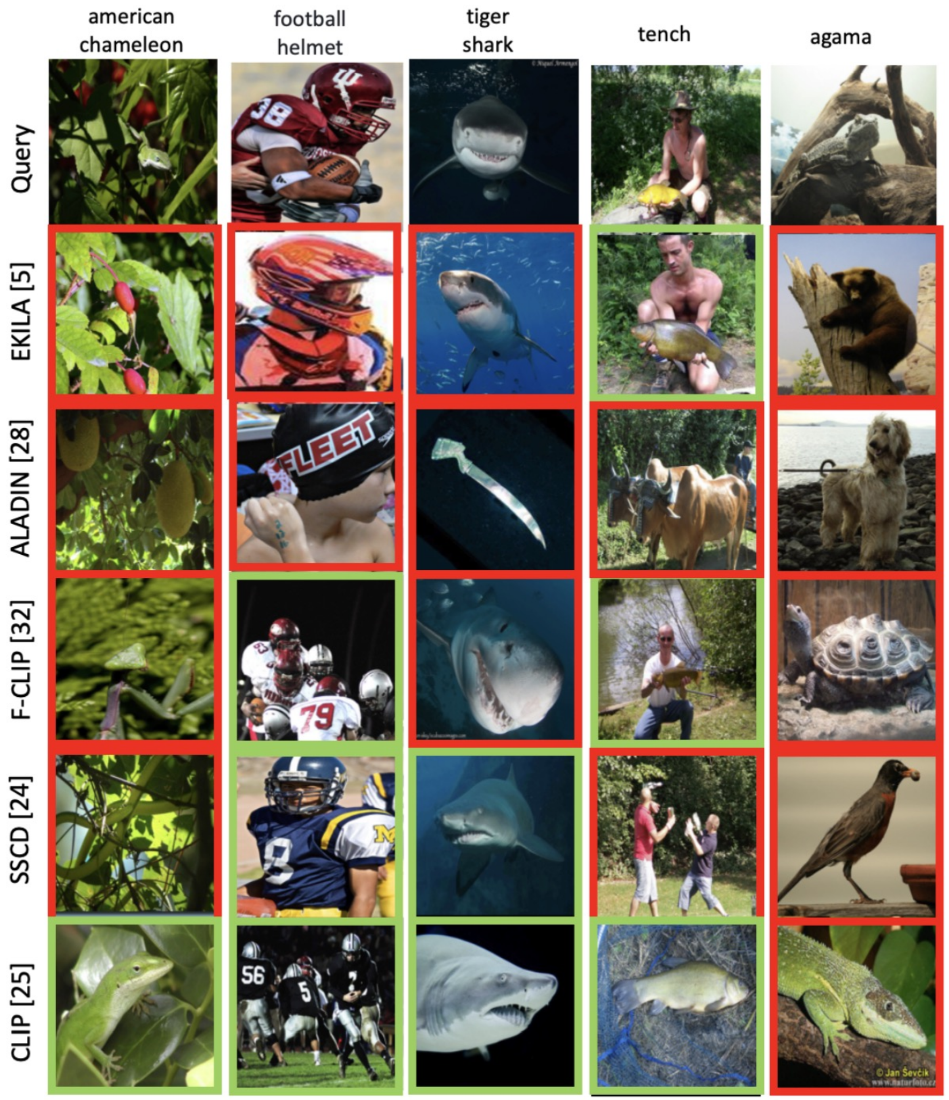
Problem: Correlation is not causation

[1] Balan, Kar, et al. "EKILA: Synthetic Media Provenance and Attribution for Generative Art." *In CVPRW* 2023.

[2] Ruta, Dan, et al. "Aladin: all layer adaptive instance normalization for fine-grained style similarity." *In ICCV*. 2021.

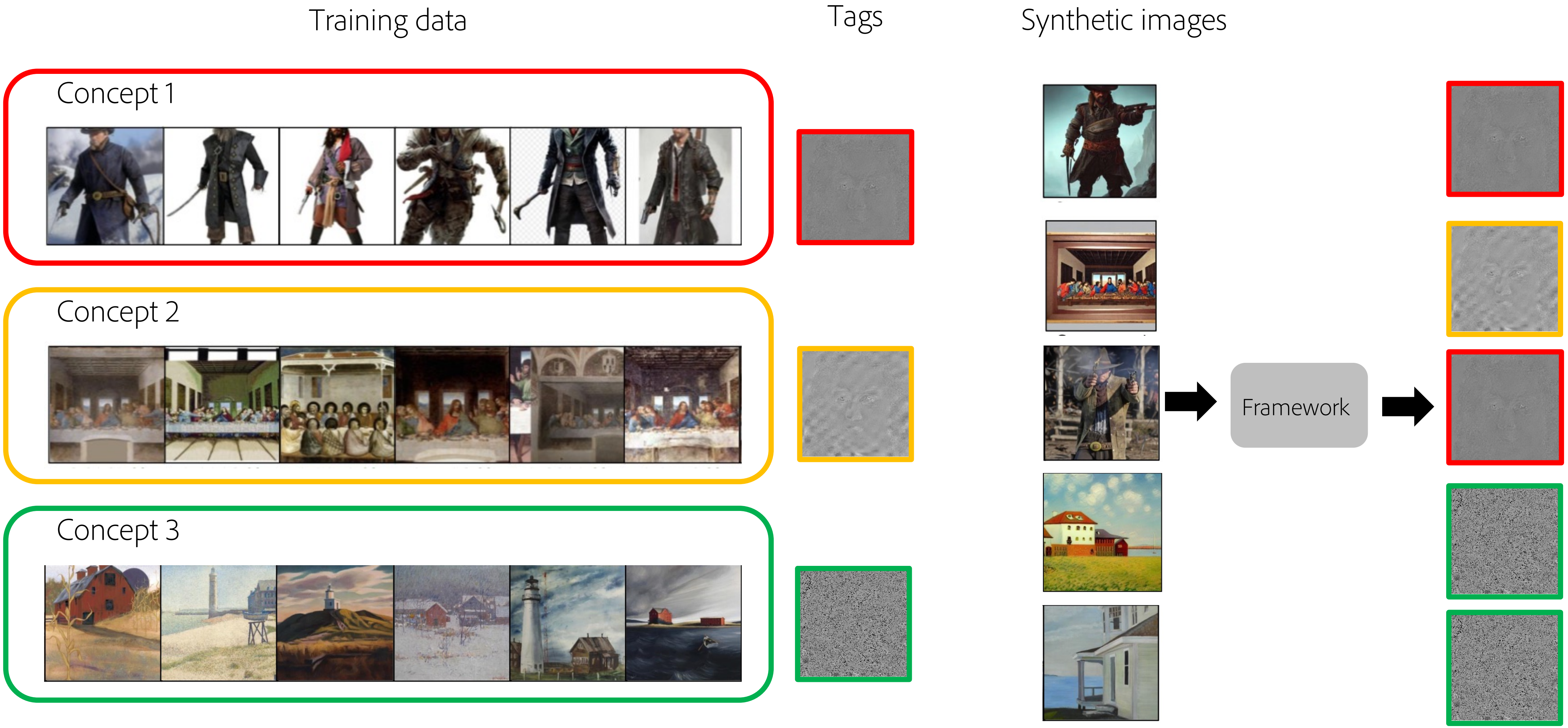
[3] Richard et al, Evaluating Data Attribution for Text-to-Image Models, submitted manuscript 2023

Prior Work Failure Cases ---> Our Motivation



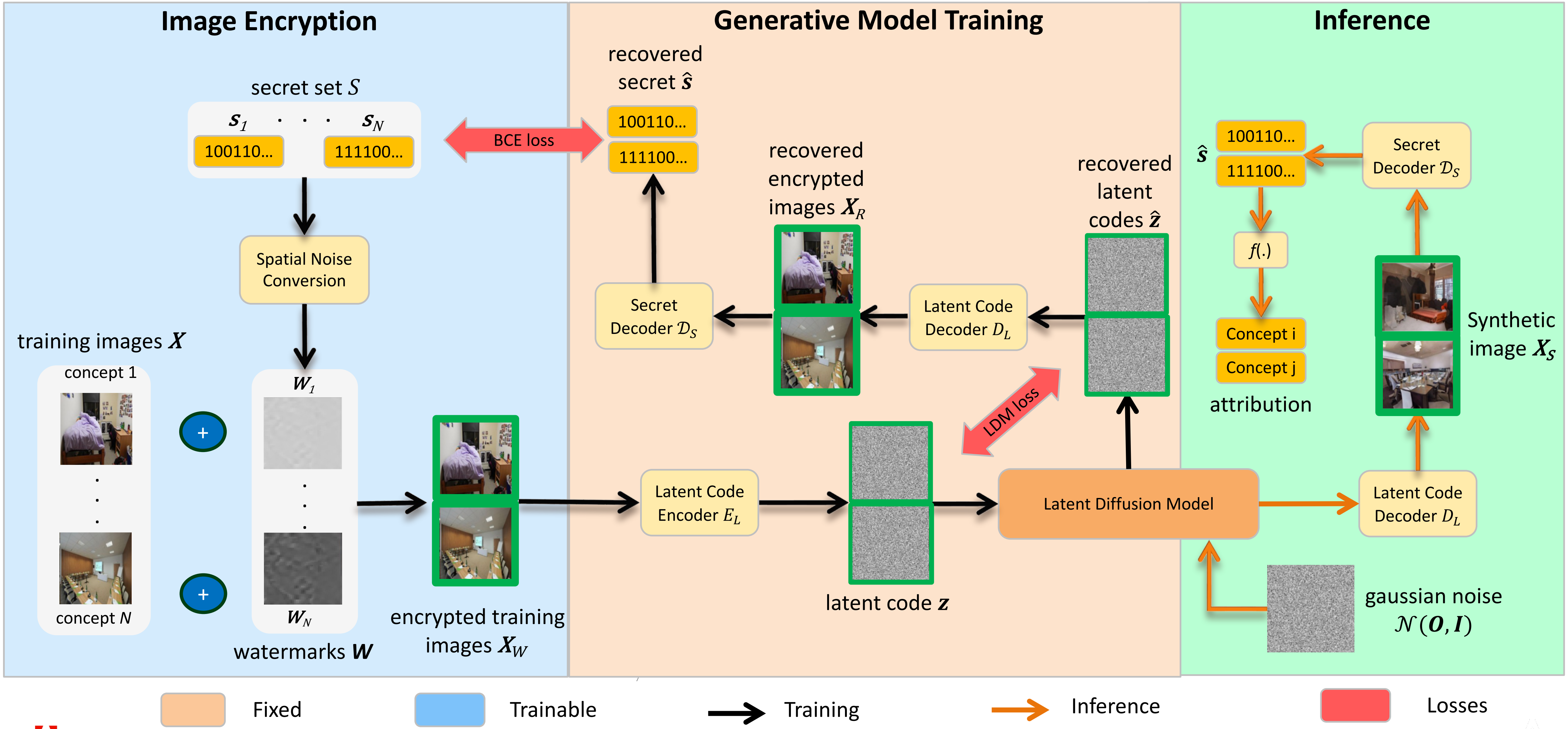
- Prior works adopt correlation matching of image embeddings
- Embedding matching works on visual similarity
- Not always results in correct concept attribution

ProMark: A causative (proactive) way



Attributing synthetic images to the training images using a tag

Framework



Adobe

Results

Single concept attribution on multiple datasets(# concepts)

Method	Str. (%)	Attribution Accuracy (%)				
		Stock (100)	LSUN (10)	Wiki-A (23)	Wiki-S (28)	ImageNet (1000)
ALADIN	-	99.86	46.27	48.95	33.25	9.25
CLIP	-	75.67	87.13	77.58	60.84	60.12
F-CLIP	-	78.49	87.39	77.23	60.43	62.83
SSCD	-	99.63	73.26	69.51	50.37	37.32
EKILA	-	99.37	70.60	51.23	37.06	38.00
ProMark	30	100	95.12	97.45	98.12	83.06
	100	**100**	**100**	**100**	**100**	**91.07**

ProMark achieves perfect attribution accuracy for multiple datasets, significantly outperforming passive works

Trade-off between image quality and attribution performance as demonstrated by watermark strength hyperparameter

Multi-concept attribution on BAM dataset with two attributes: media (7 concepts) and content (8 concepts)

Method	Str. (%)	Attribution Accuracy (%)		
		Media (7)	Content (8)	Combined (7 x 8)
ALADIN	-	42.16	41.25	34.97
CLIP	-	46.71	45.12	42.36
F-CLIP	-	52.12	51.56	46.23
SSCD	-	47.06	46.09	40.61
EKILA	-	43.72	43.58	37.09
ProMark (single)	30	-	-	**97.73**
ProMark (multi)	30	91.33	89.21	84.66
	100	**95.61**	**93.31**	90.12

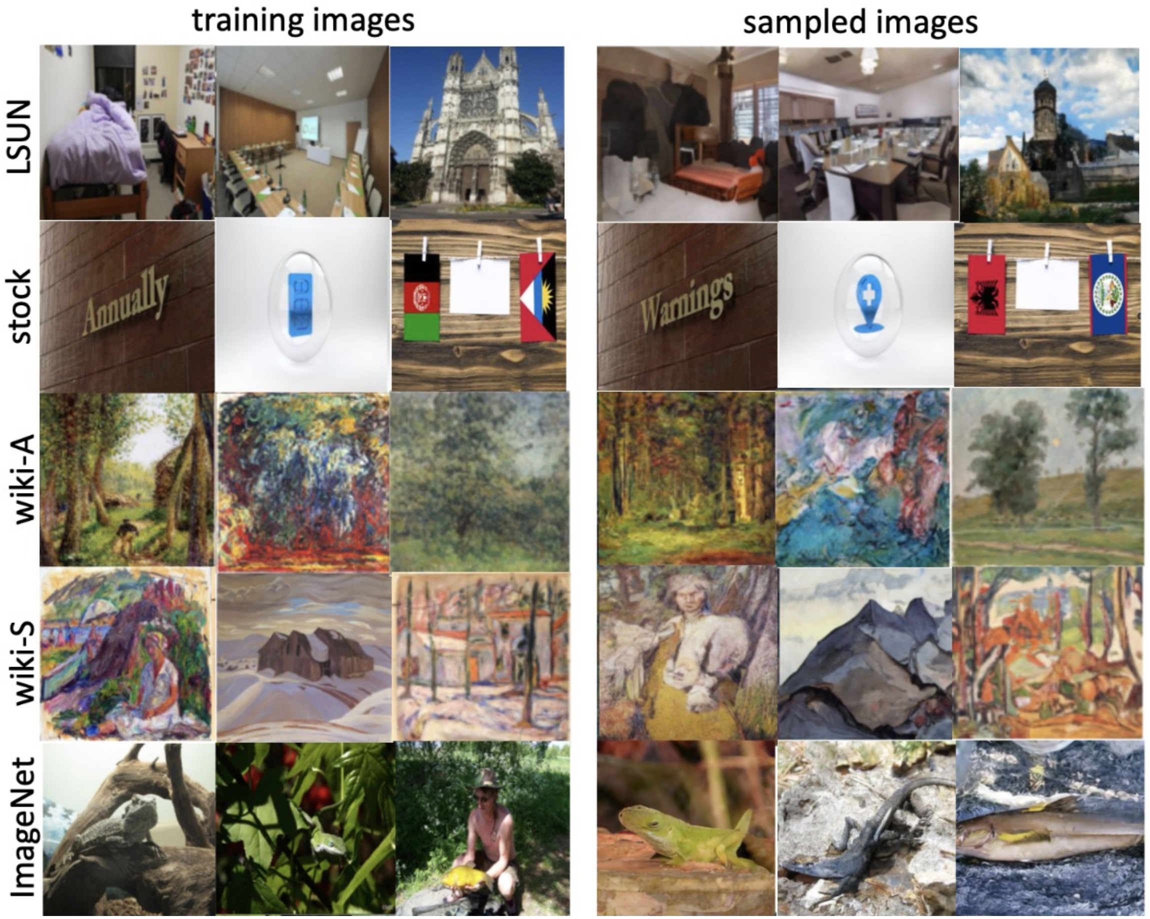
Better multi/single combined attribution accuracy

Better individual concept attribution accuracy

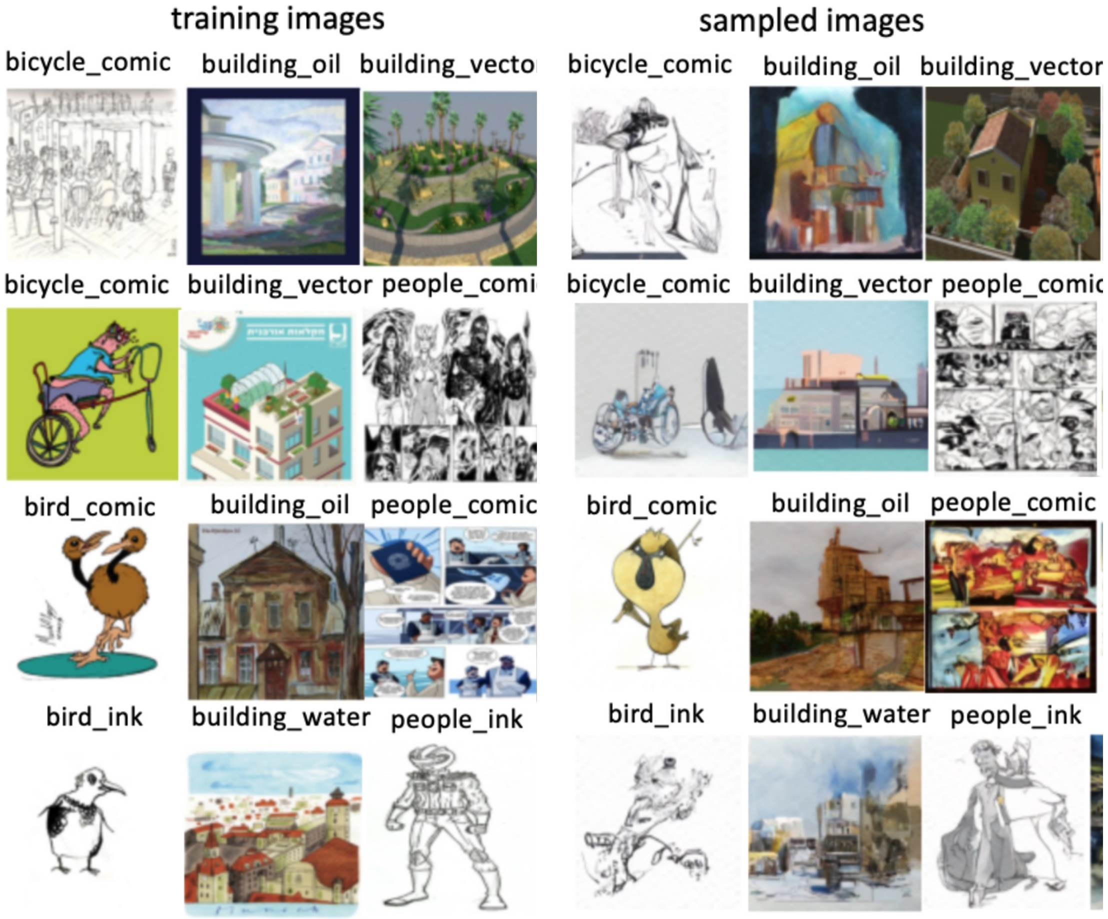
Single concept approach not scalable: single (7 x 8 concepts) vs multiple (7 + 8 concepts)

Visualization

Single concept

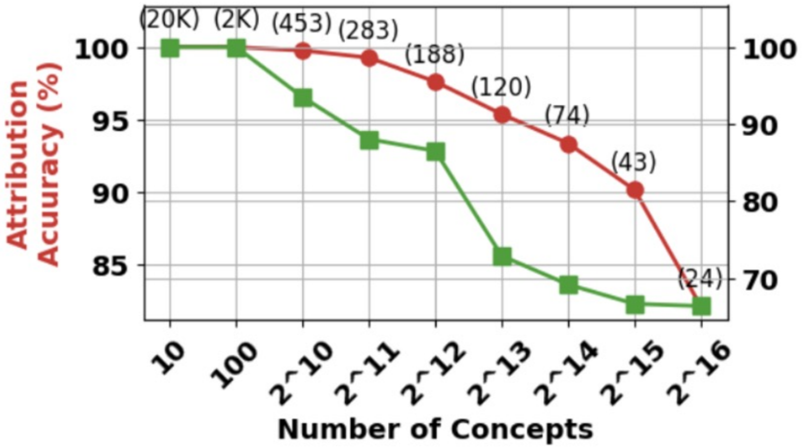


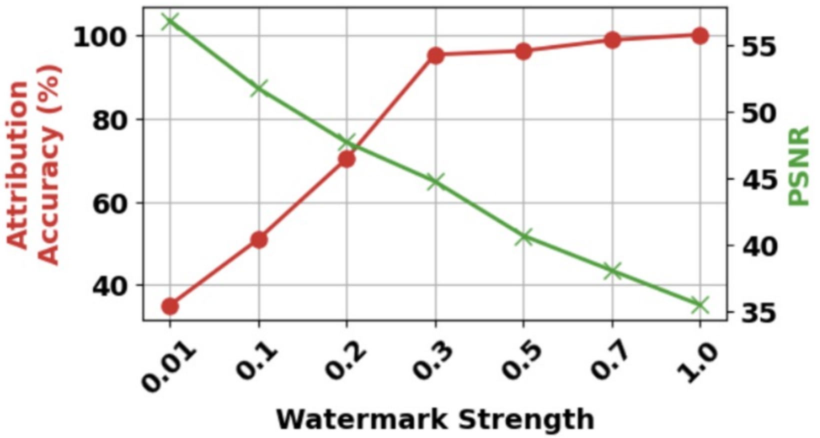
Multiple concept

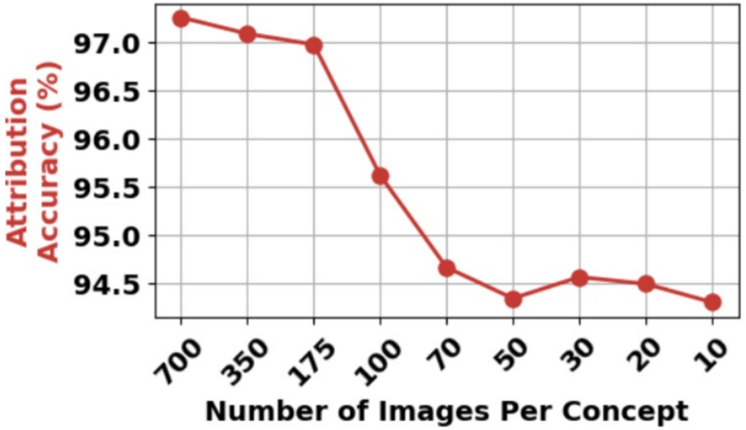


1:1 correspondence between training and newly sampled image.

Ablation Studies







Embed as many as 2^{16} concepts

Number of secrets ⬆

Attribution accuracy ⬇

Bit-secret accuracy ⬇

Quality-performance trade-off

Watermark strength ⬆

Attribution accuracy ⬆

PSNR ⬇

ProMark learns attribution with as low as 10 images per concept

Conclusion

- ProMark is a proactive watermarking-based approach for concept attribution in generative AI models.
- It uses imperceptible watermarks embedded into training images.
- ProMark attributes as many as 2^{16} unique training-data concepts.
- It achieves higher attribution accuracy compared to correlation-based passive attribution methods.
- ProMark can be used for both single-concept and multi-concept attribution.

Thank You!!

Paper

Poster: Thursday morning, #: 111

Adobe

MICHIGAN STATE UNIVERSITY

UNIVERSITY OF SURREY