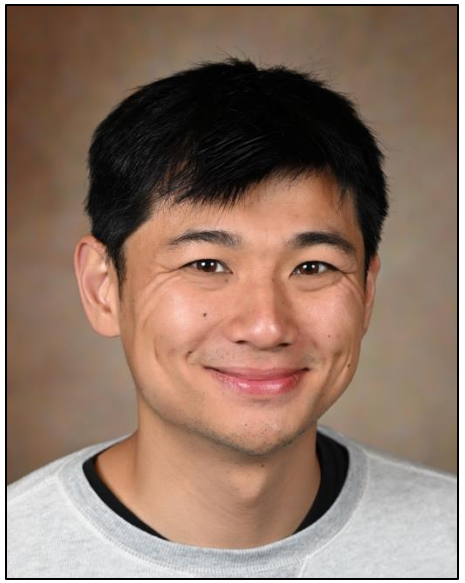
Data Attribution for Text-to-Image Models by Unlearning Synthesized Images

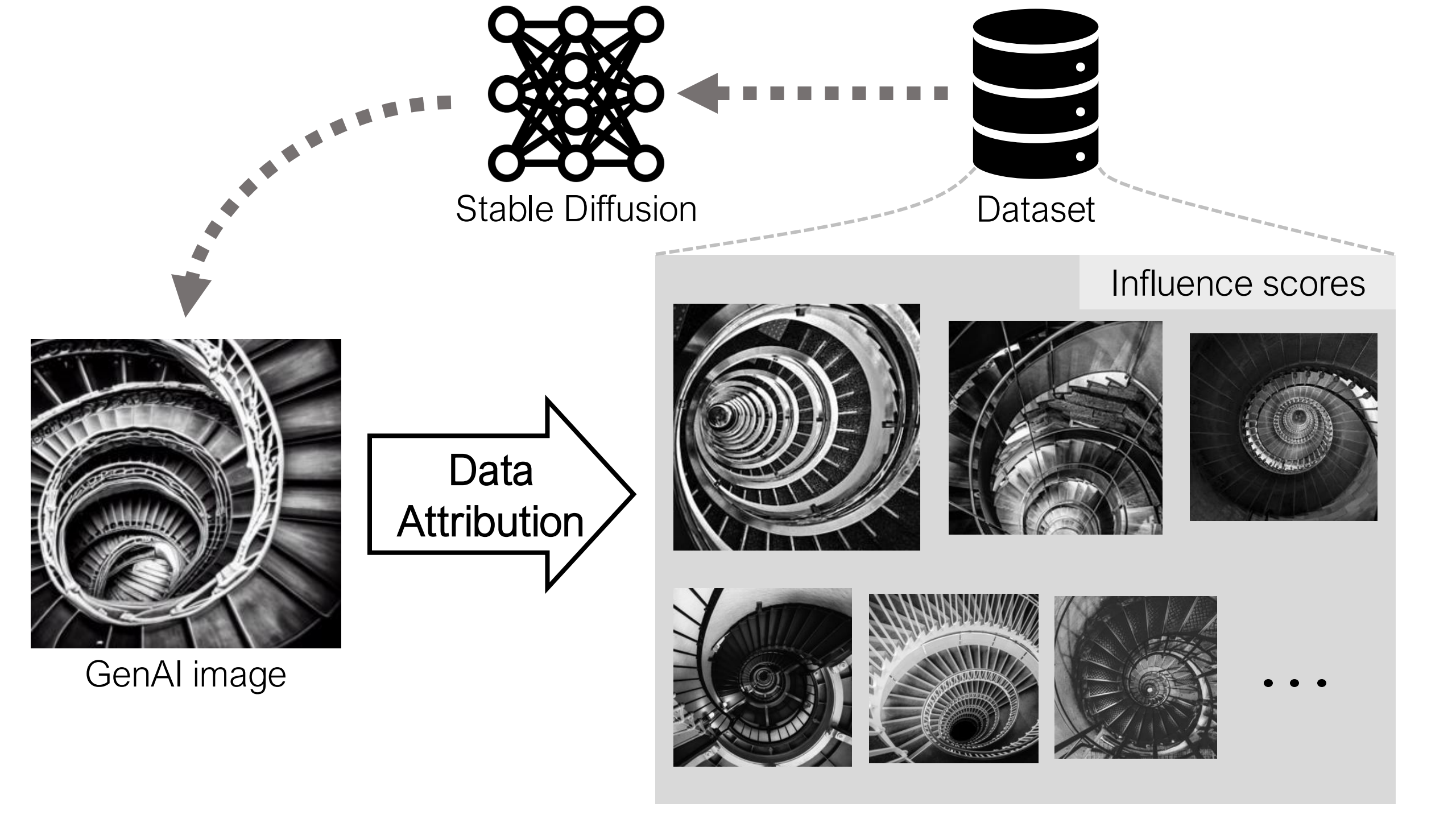


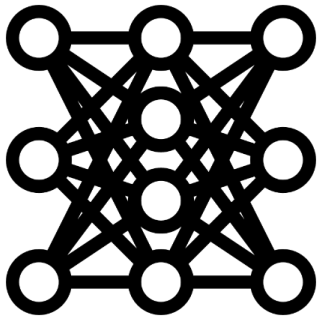




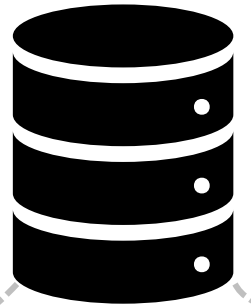
Sheng-Yu Wang Aaron Hertzmann Alexei A. Efros Jun-Yan Zhu Richard Zhang

In NeurIPS, 2024.


Stable Diffusion
Dataset
Influence scores
GenAI image
Data Attribution
. . .



Stable Diffusion



Dataset

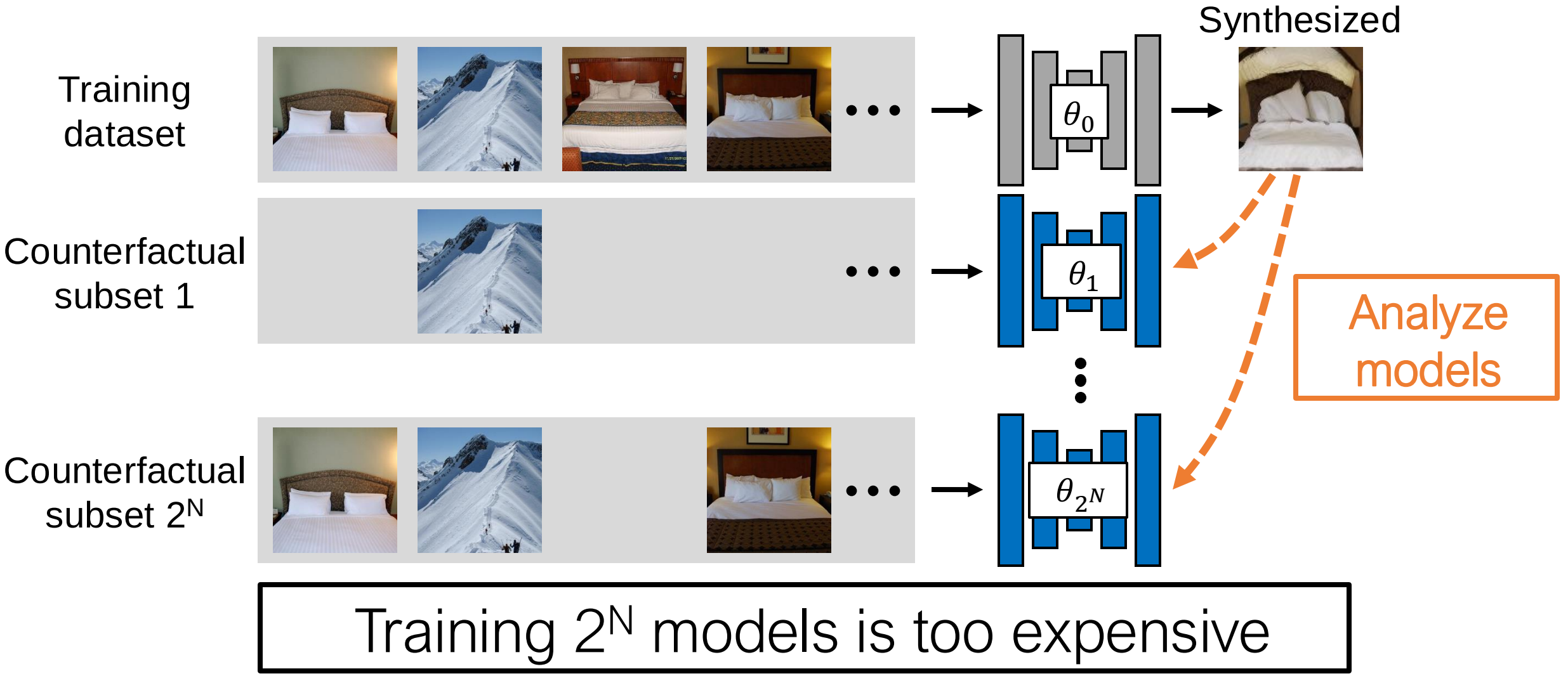
Challenge: ground truth influence is unknown…
Must intervene in the training process

GenAI image

Data Attribution

. . .

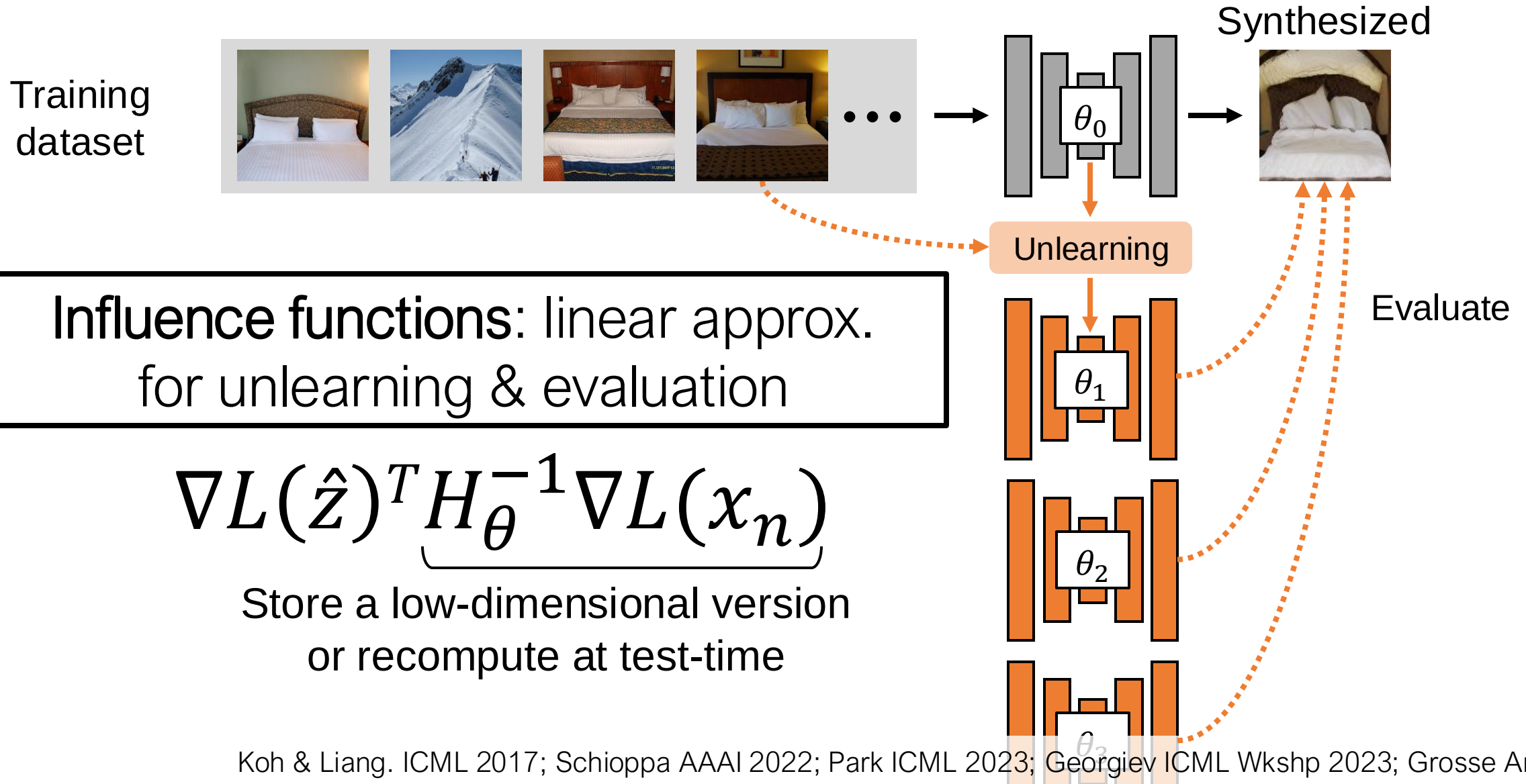
Random subsets



Training 2^N models is too expensive

c.f. Feldman & Zhang. What Neural Networks Memorize and Why. NeurIPS 2020.

Leave-one-out



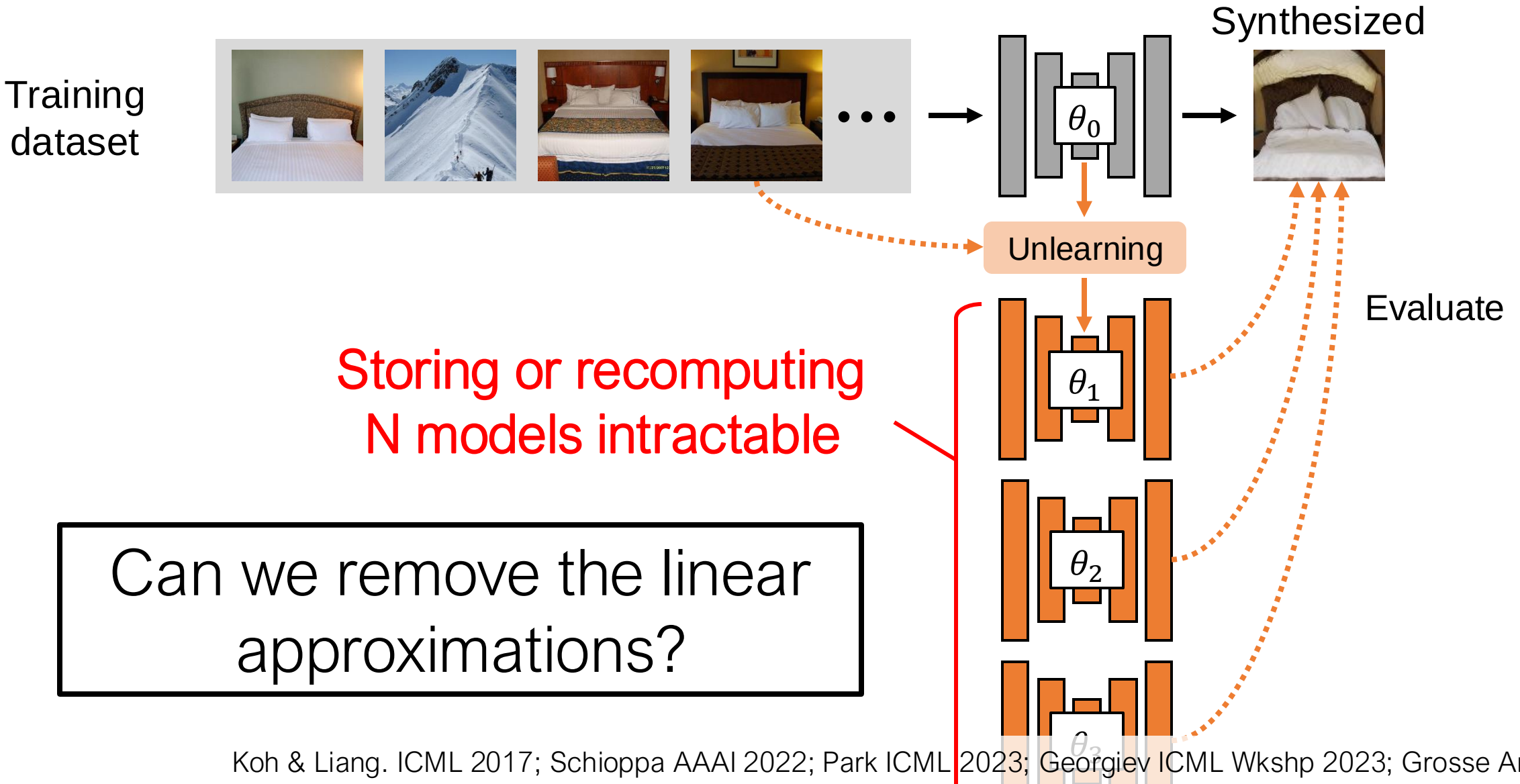
Influence functions: linear approx. for unlearning & evaluation

$$\nabla L(\hat{z})^T \underbrace{H_\theta^{-1}} \nabla L(x_n)$$

Store a low-dimensional version or recompute at test-time

Koh & Liang. ICML 2017; Schioppa AAAI 2022; Park ICML 2023; Georgiev ICML Wkshp 2023; Grosse ArXiv 2023.

Leave-one-out

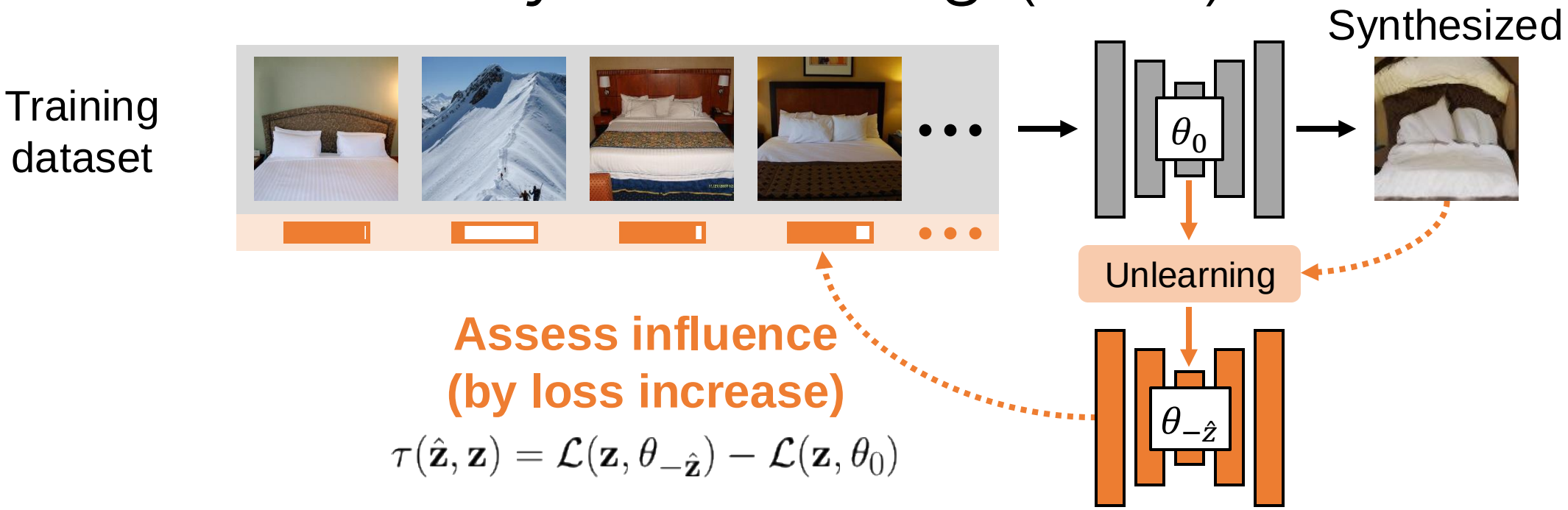


Storing or recomputing N models intractable

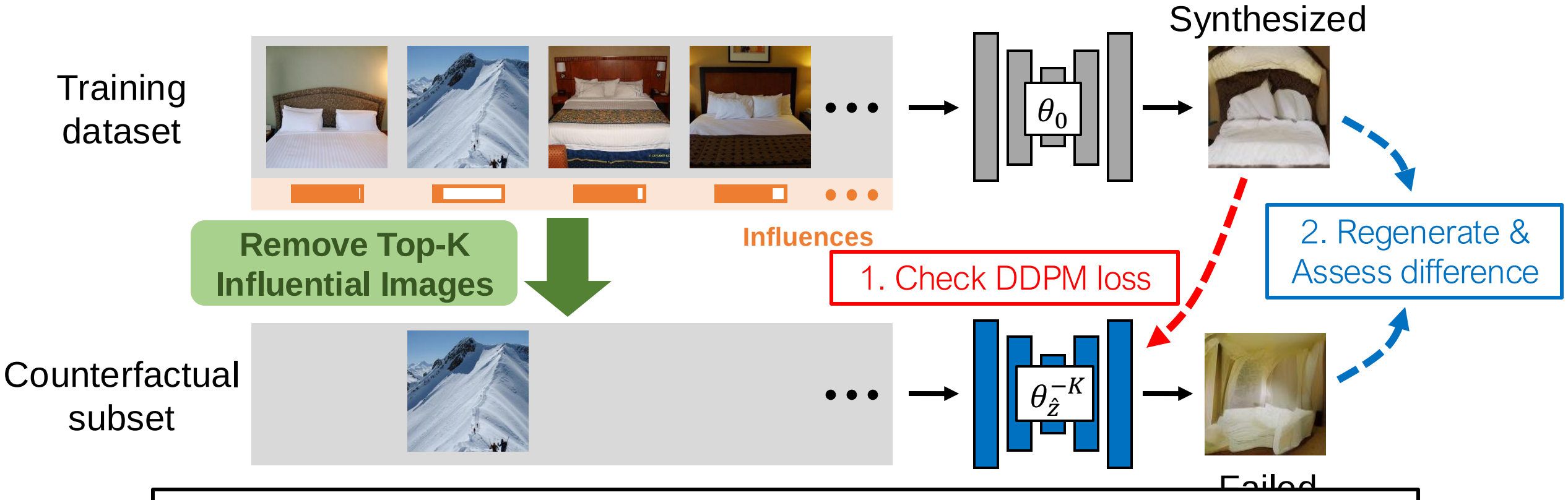
Can we remove the linear approximations?

Koh & Liang. ICML 2017; Schioppa AAAI 2022; Park ICML 2023; Georgiev ICML Wkshp 2023; Grosse ArXiv 2023.

Attribution by Unlearning (AbU)



Counterfactual evaluation



If critical training images are identified, removing them should destroy the generation

c.f. K. Georgiev, et al. How Training Data Guides Diffusion Models. In ArXiv, 2023.

MS-COCO results

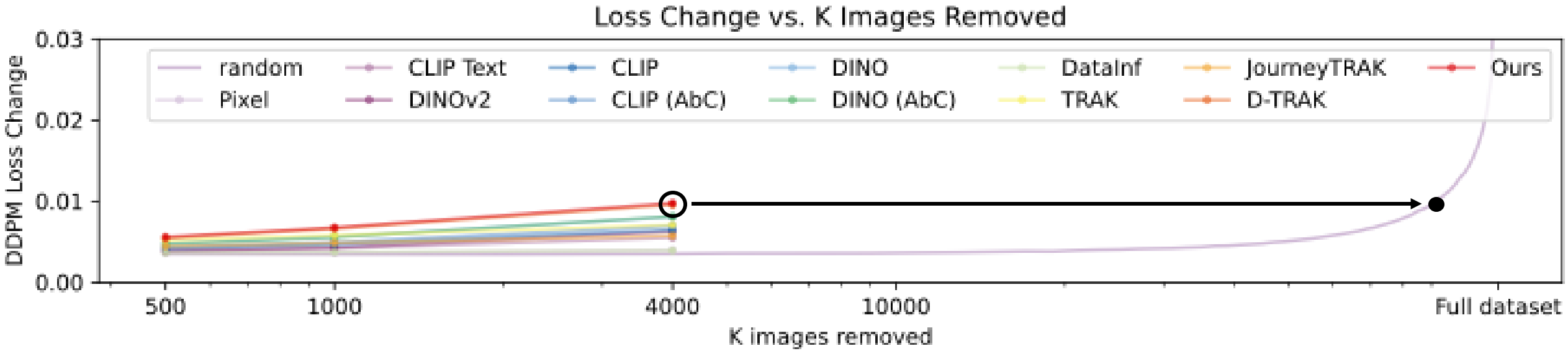
"A bus traveling on a freeway next to other traffic."

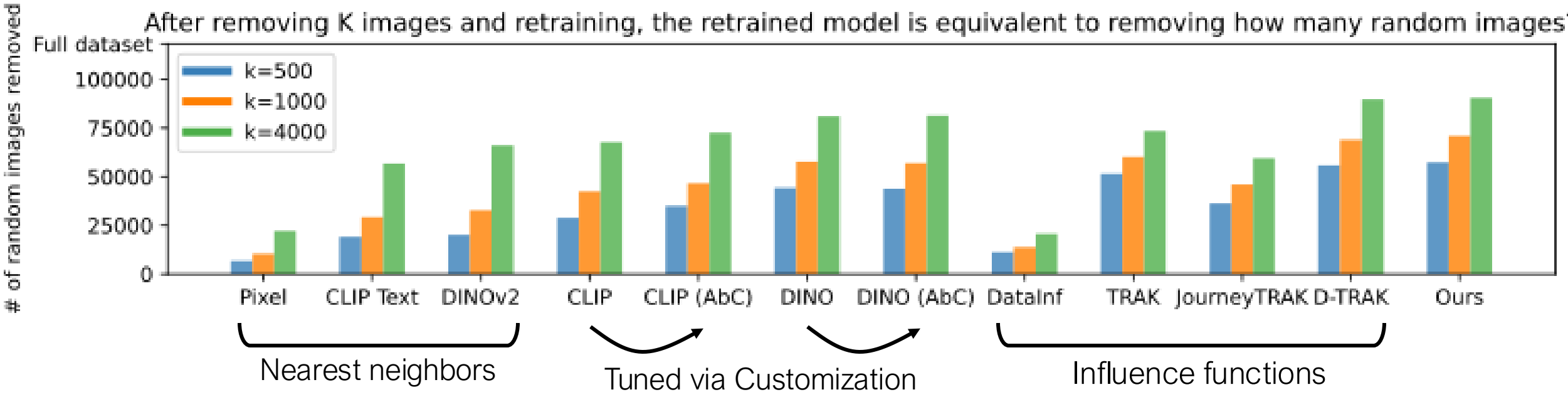
Attribution results

Effective removal

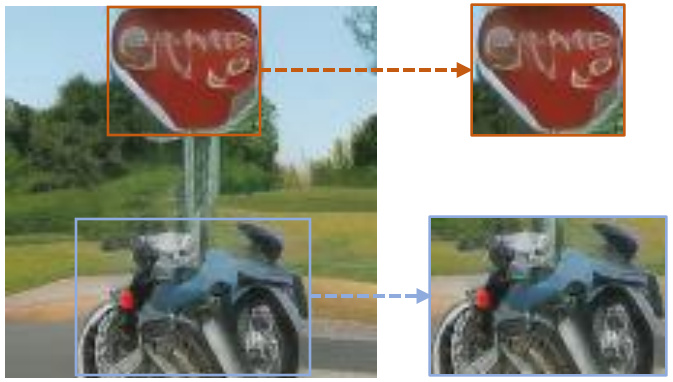
Remove K=500
(0.4% of dataset)

c.f. K. Georgiev, et al. How Training Data Guides Diffusion Models. In ArXiv, 2023.
c.f. X. Zheng, et al. Intriguing Properties of Data Attribution on Diffusion Models. In ICLR, 2024.


Loss Change vs. K Images Removed
DDPM Loss Change
0.03
0.02
0.01
0.00
random
Pixel
CLIP Text
DINOv2
CLIP
CLIP (AbC)
DINO
DINO (AbC)
DataInf
TRAK
JourneyTRAK
D-TRAK
Ours
500
1000
4000
10000
Full dataset
K images removed


After removing K images and retraining, the retrained model is equivalent to removing how many random images?
# of random images removed
Full dataset
100000
75000
50000
25000
0
k=500
k=1000
k=4000
Pixel
CLIP Text
DINOv2
CLIP
CLIP (AbC)
DINO
DINO (AbC)
DataInf
TRAK
JourneyTRAK
D-TRAK
Ours
Nearest neighbors
Tuned via Customization
Influence functions

Local attribution



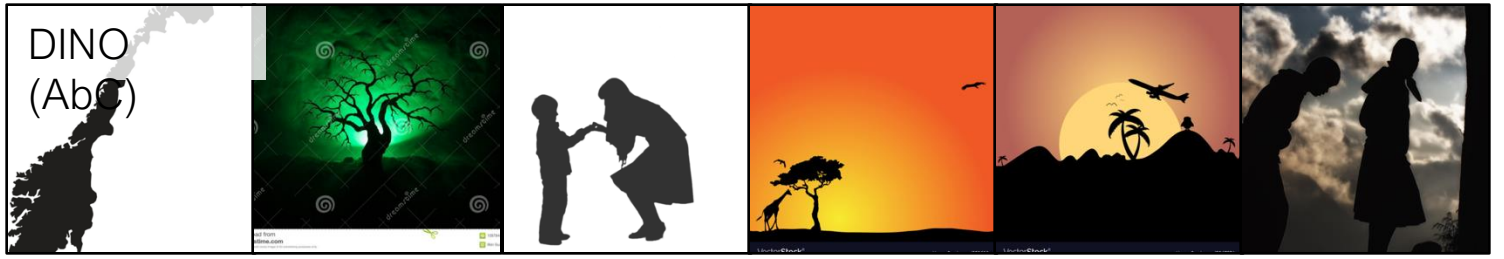
"A motorcycle and a stop sign."

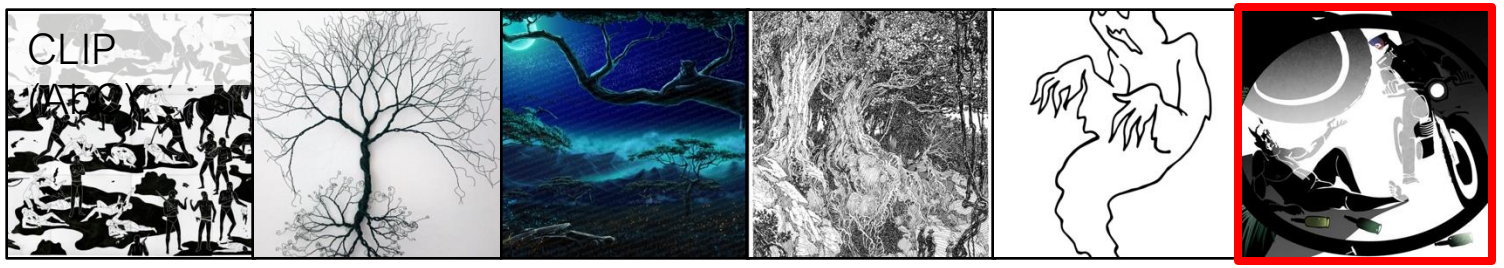
Cropped Queries

Attributed training images

Customized Model Benchmark

"A picture of tree in the style of V* art"





Thank you!

peterwang512.github.io/AttributeByUnlearning