Evaluating Data Attribution for Text-to-Image Models

By Max Zakharchenko

Idea

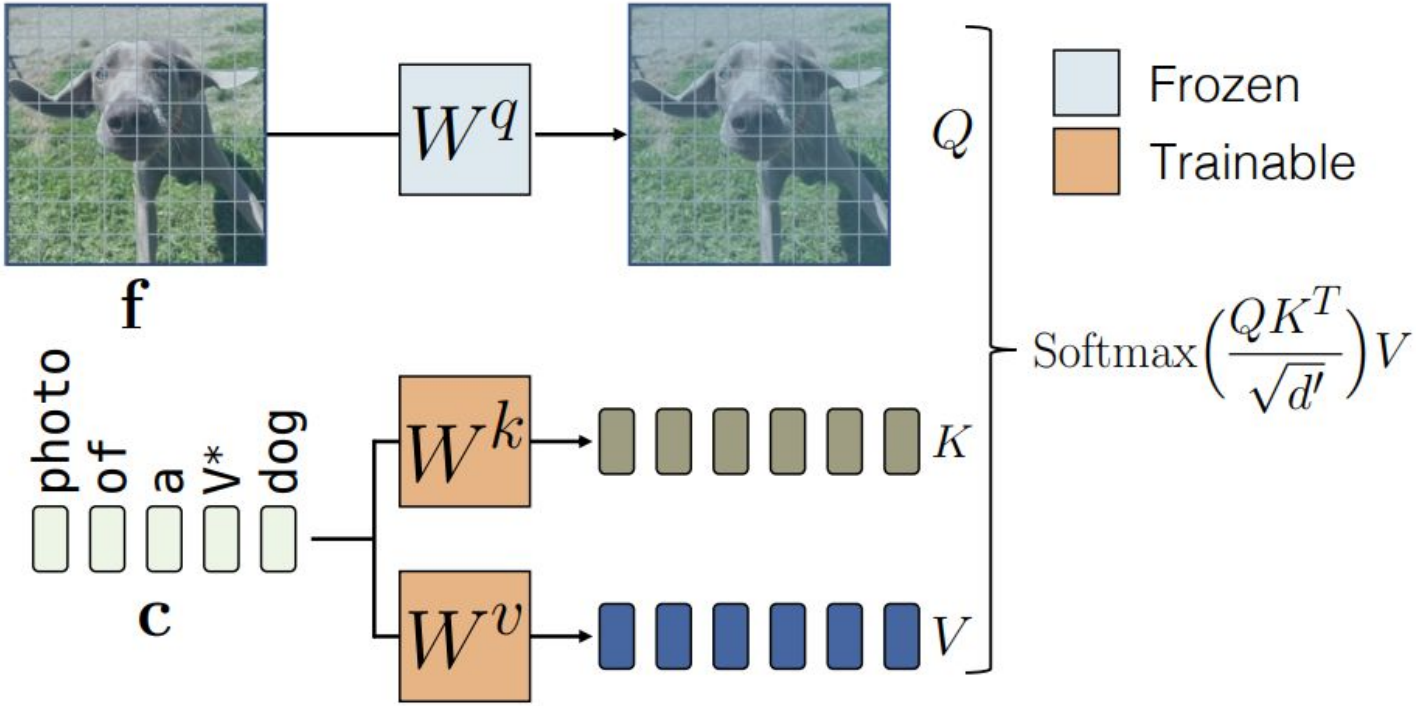
We want to know which images contribute to output

Because of copyright and ownership of the training images, understanding the interplay between training data and generative model outputs has become increasingly necessary, both for scientific progress, as well as for practical or legal reasons.

Custom diffusion TLDR

- Very efficient finetune (6 minutes on 2 A100) with only 75 MB of additional weights
- Special token V^* for new concepts
- Only train key and value matrices in only cross-attention
- Saved only row-rank approximation

Custom diffusion TLDR



Dataset. Object-centric models. Images

- Select 10 images per 693 ImageNet-1K classes
- Build 2 datasets
  - Seen classes (5930 images)
  - Unseen classes (ImageNet-100 classes) for out-of-distribution testing

Dataset. Object-centric models. Prompts

- Training prompt
  - `V* cat`
- Chat-GPT (tends to realism)
  - `The` **`V* cat`** `groomed itself meticulously`
- Medium Prompt
  - `A <medium> of` **`V* cat`**
  - `<medium>` is a sample from `watercolor painting, tattoo, digital art`

Dataset. Object-centric models

Property		Object-centric				
		Imagenet-Seen			Unseen	Total
		train	val	test	test	
Object classes		593	593	593	100*	693
Training images		4744	593	593	1000	6930
Avg images/model		1	1	1	1	1
Total models		4744	593	593	1000	6930
Prompts	ChatGPT†	15	6	10	10	–
	Procedural	40	6	10‡	10‡	50
Samples	ChatGPT	284,640	14,232	23,720	40,000	362,592
	Procedural	759,040	14,232	23,720	40,000	836,992
	Total	1,043,680	28,464	47,440	80,000	1,199,584

Dataset. Artistic-style models.

- BAM-FG and Artchive datasets
- Procedural `A picture in the style of` **`V*`** **`art`** **training** prompts
- Chat-GPT's inference prompts
  - Sample 50 painting captions
  - `The magic of the forest in the style of` **`V*`** **`art`**
- Procedural inference prompts
  - 40 different objects, such as flowers and rivers
  - `A picture of in the style of` **`V*`** **`art`** for BAM-FG
  - `A painting ..` for Artchive

Dataset. Artistic-style models.

Property		Artistic styles				
		BAM-FG			Artchive	Total
		train	val	test	test	
Object classes		–	–	–	–	–
Training images		78,086	1837	1692	3081	84,696
Avg images/model		7.36	7.35	6.77	12.08	7.45
Total models		10,607	250	250	255	11,362
Prompts	ChatGPT†	40	6	10	10	50
	Procedural	30	6	10‡	10‡	40
Samples	ChatGPT	1,697,120	6,000	10,000	10,200	1,723,320
	Procedural	1,272,840	6,000	10,000	10,200	1,299,040
	Total	2,969,960	12,000	20,000	20,400	3,022,360

Dataset. Summary.

We have *N* models

- X_k as training dataset
- X_k as influenced synthetic images dataset (inference dataset)

Goal is to predict X_k from X_k

Evaluating existing features

Define feature extraction *F* (CLIP, DINO) such that

$$\mathrm{sim}\left(F(x),\ F(x)\right) > \mathrm{sim}\left(F(x),\ F(x)\right)$$

“assessing visual similarity is not equivalent to attributing data influence”

Learning features for attribution

Define feature extractions *F* and *F* and some light mapping functions *h* and *h*

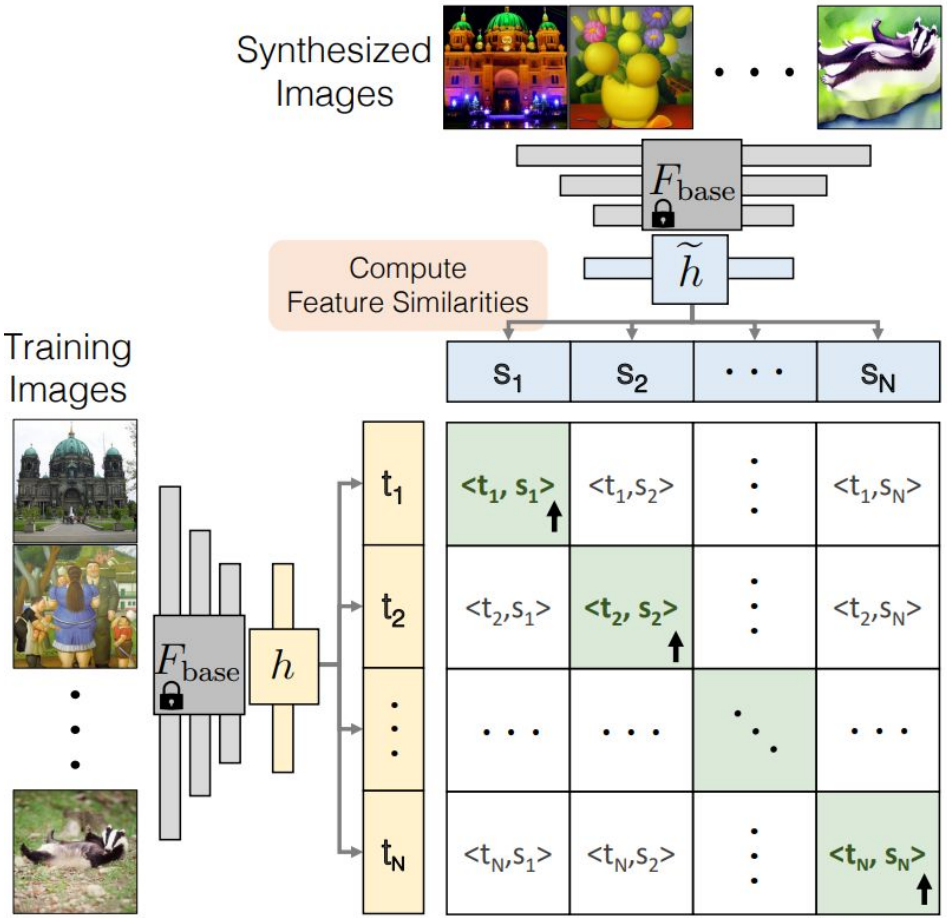
$$F_{\text{base}} - \text{pretrained encoder}$$

$$F = h \circ F_{\text{base}} \qquad F = h \circ F_{\text{base}}$$

Get NT-Xent (the normalized temperature-scaled cross entropy loss)

$$\mathcal{L}^i_{\text{cont}} = -\left(\log \frac{\exp(\mathbf{t}_i^\top \mathbf{s}_i / \upsilon)}{\sum_j \exp(\mathbf{t}_i^\top \mathbf{s}_j / \upsilon)} + \log \frac{\exp(\mathbf{t}_i^\top \mathbf{s}_i / \upsilon)}{\sum_j \exp(\mathbf{t}_j^\top \mathbf{s}_i / \upsilon)} \right)$$

v is 1 in training



Contrastive Learning Across Two Views

Learning features for attribution

Extract probabilistic influence $P\left(x|x\right)$

Define loss

$$\min_{P} \mathbb{E}_{x}\left[\mathcal{D}_{\mathrm{KL}}\mathcal{S}(x;x) \parallel P(x|x)\right] \qquad \mathcal{S}(x;x) = \frac{1}{|X|}\mathbb{I}\left[x \in X\right]$$

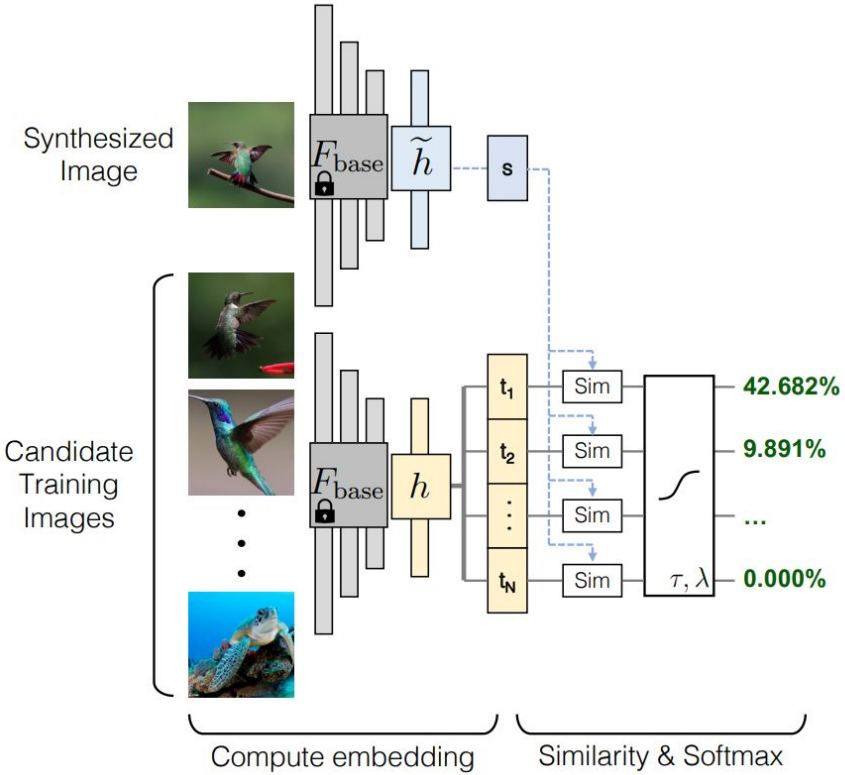
Learning features for attribution

Merging similarity and probabilistic influence

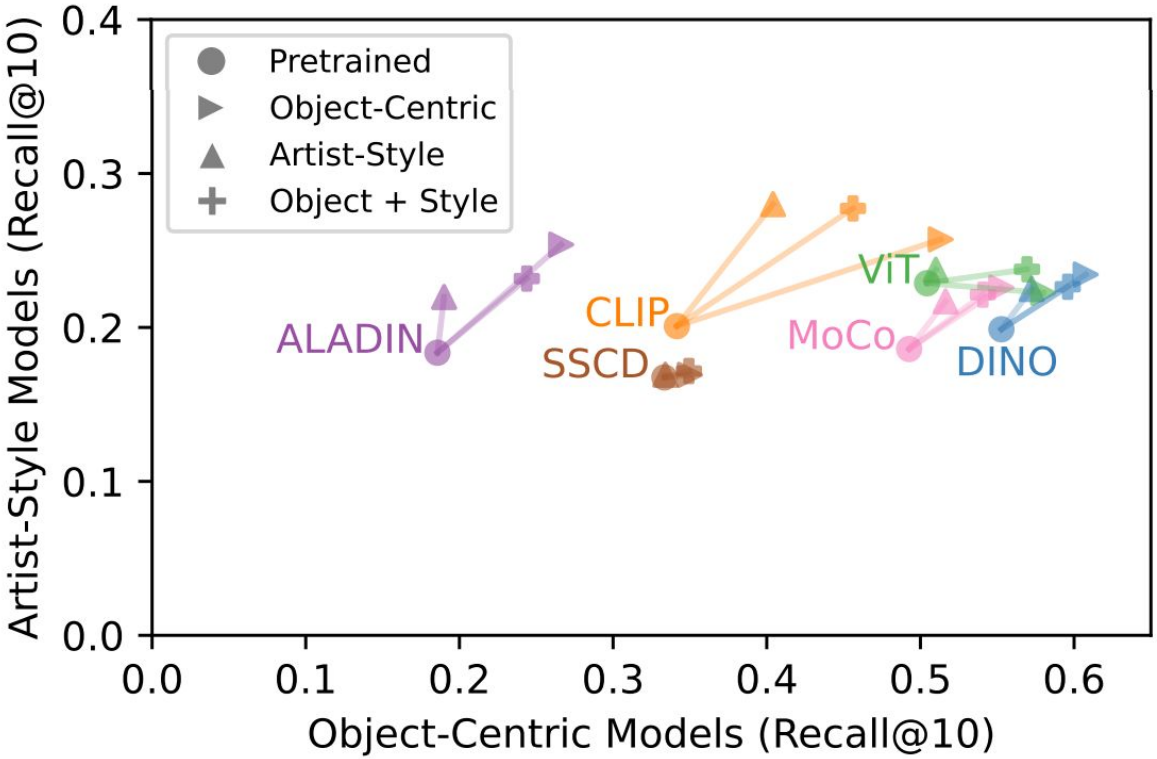
$$P_{\tau,\lambda}(x|x) = \frac{\mathrm{ReLU}(\exp(\frac{s-s_0}{t}) - \lambda)}{\sum_j \mathrm{ReLU}(\exp(\frac{s_j-s_0}{t}) - \lambda)}$$

where *s* - similarity. τ = 1

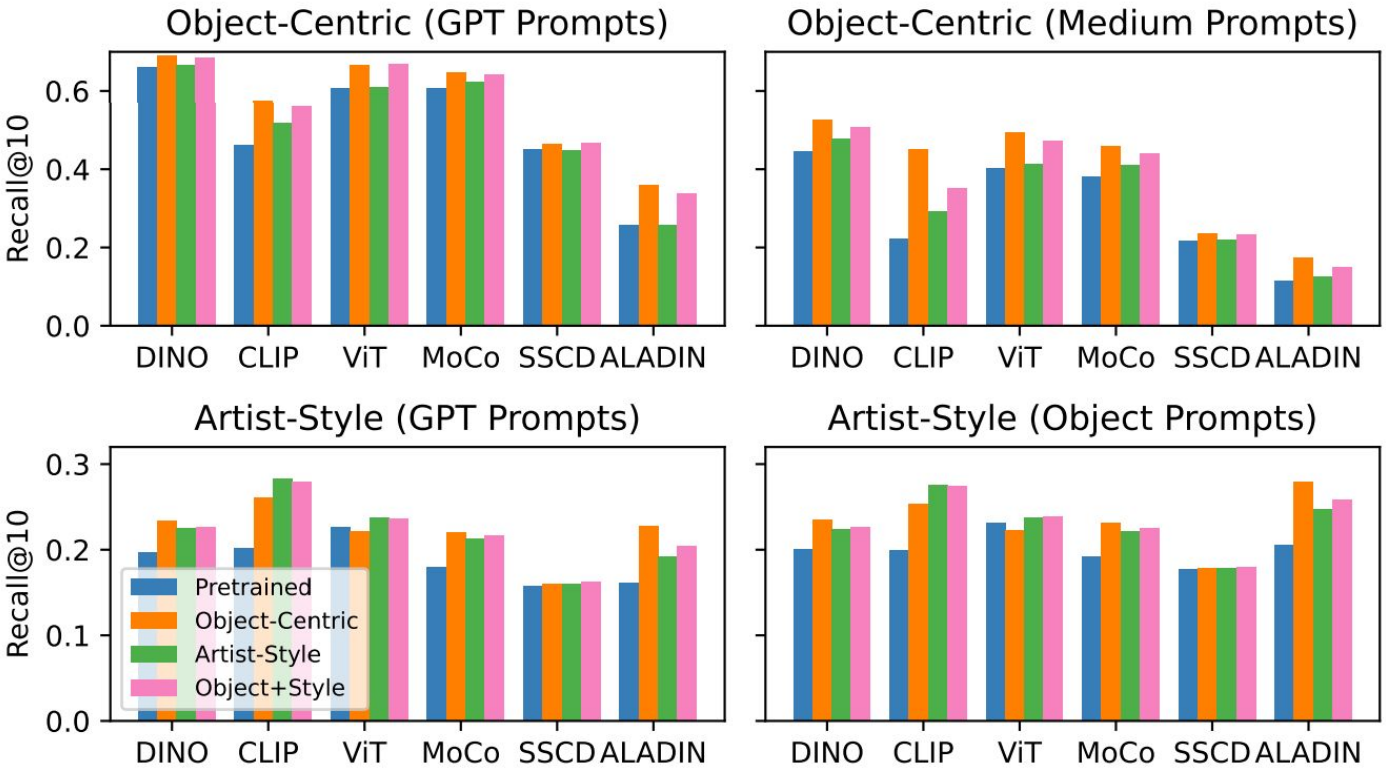
Learning features for attribution



Experiments

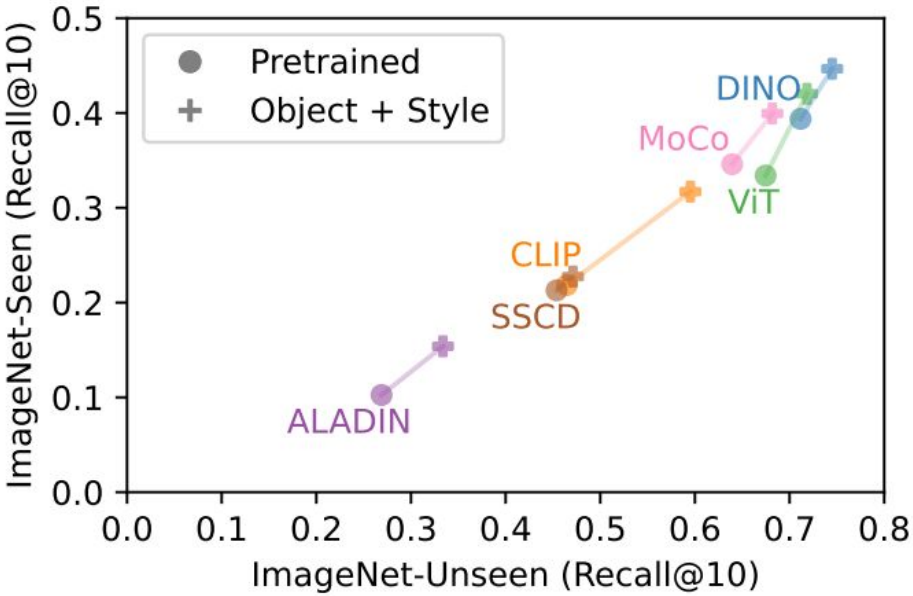


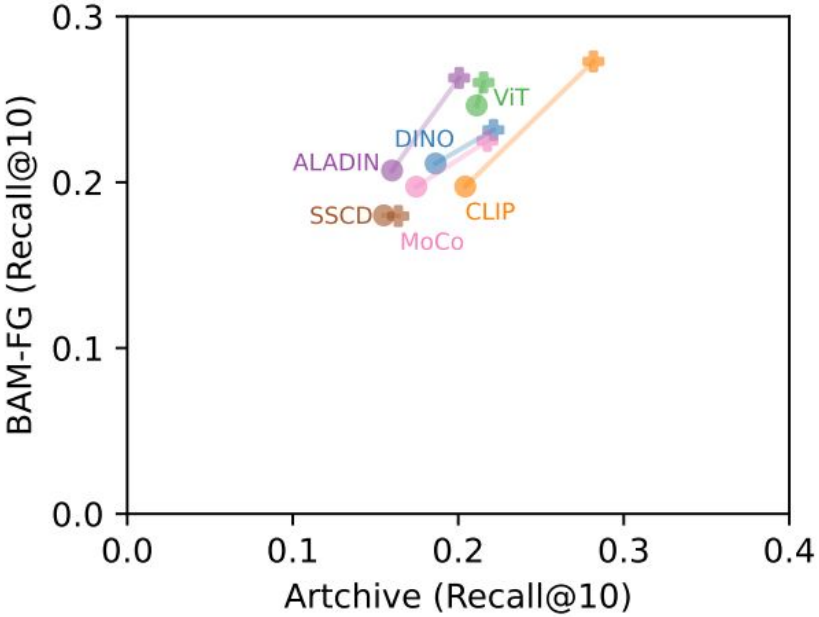
Experiments



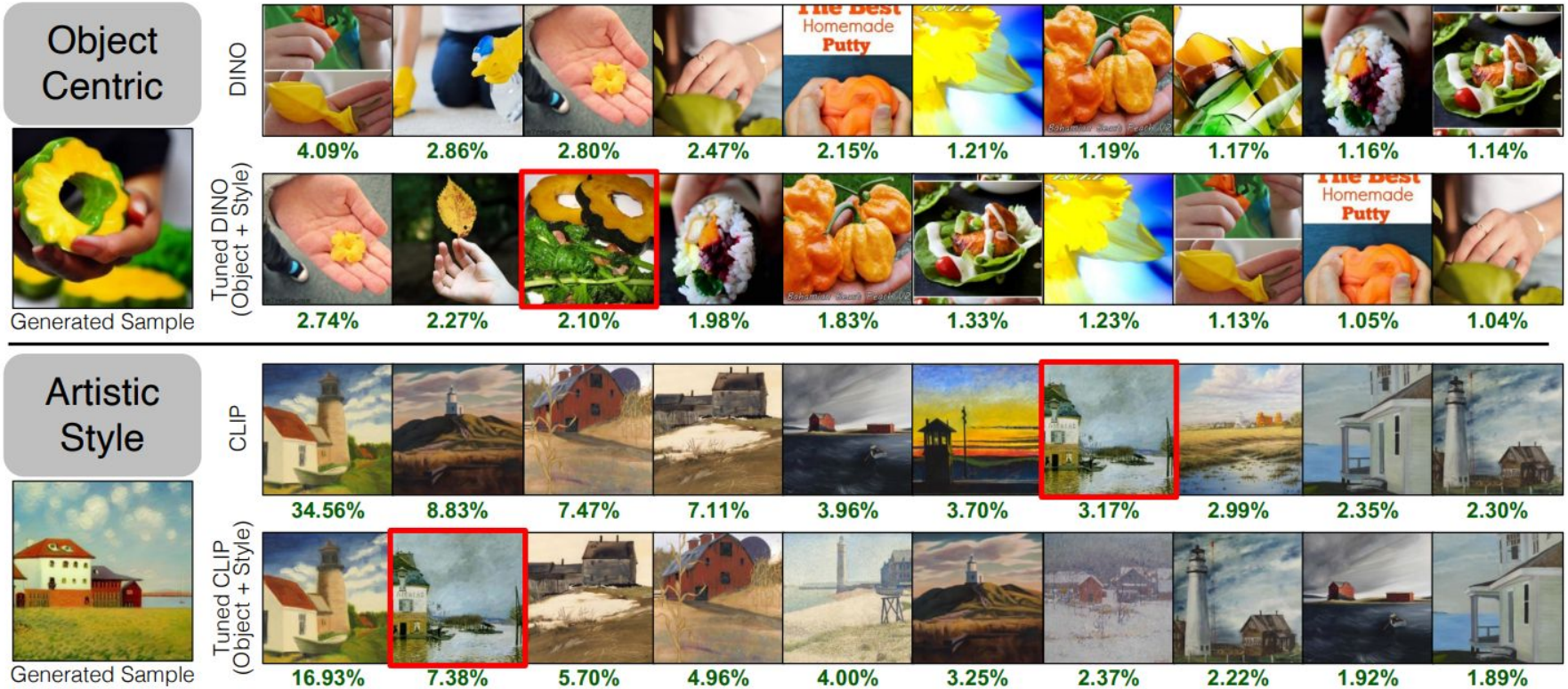
Experiments

No overfitting!

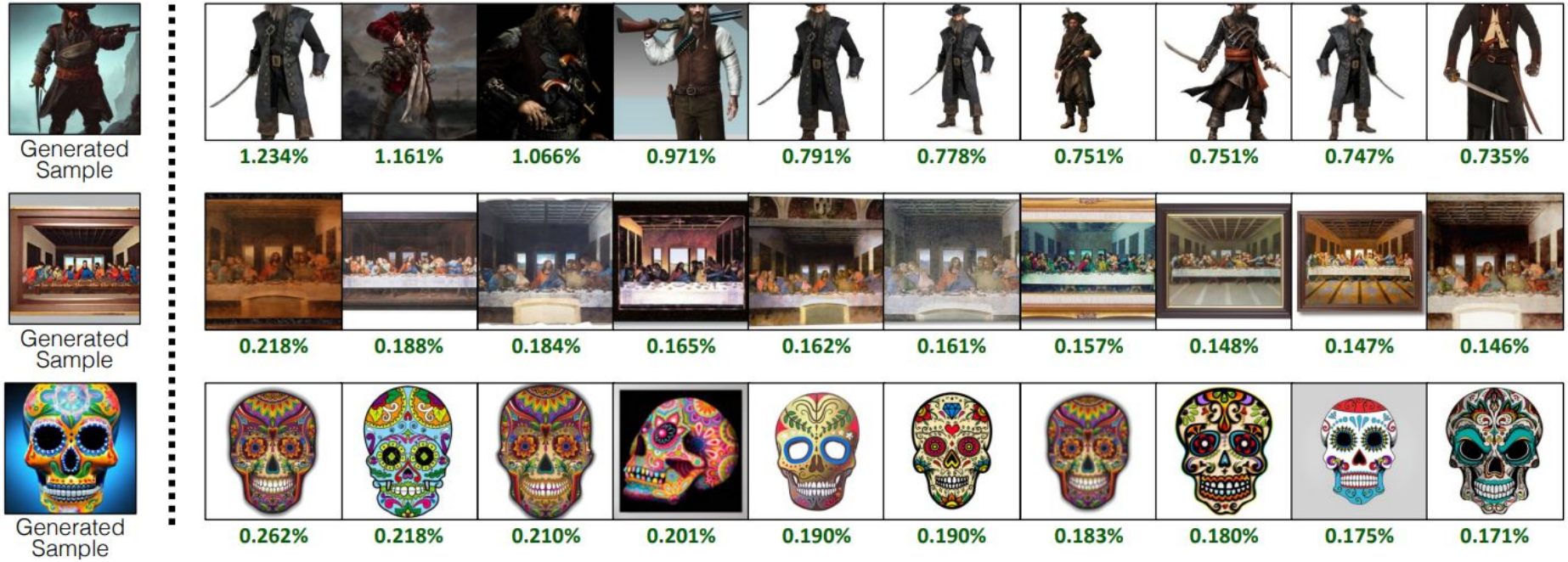




Experiments

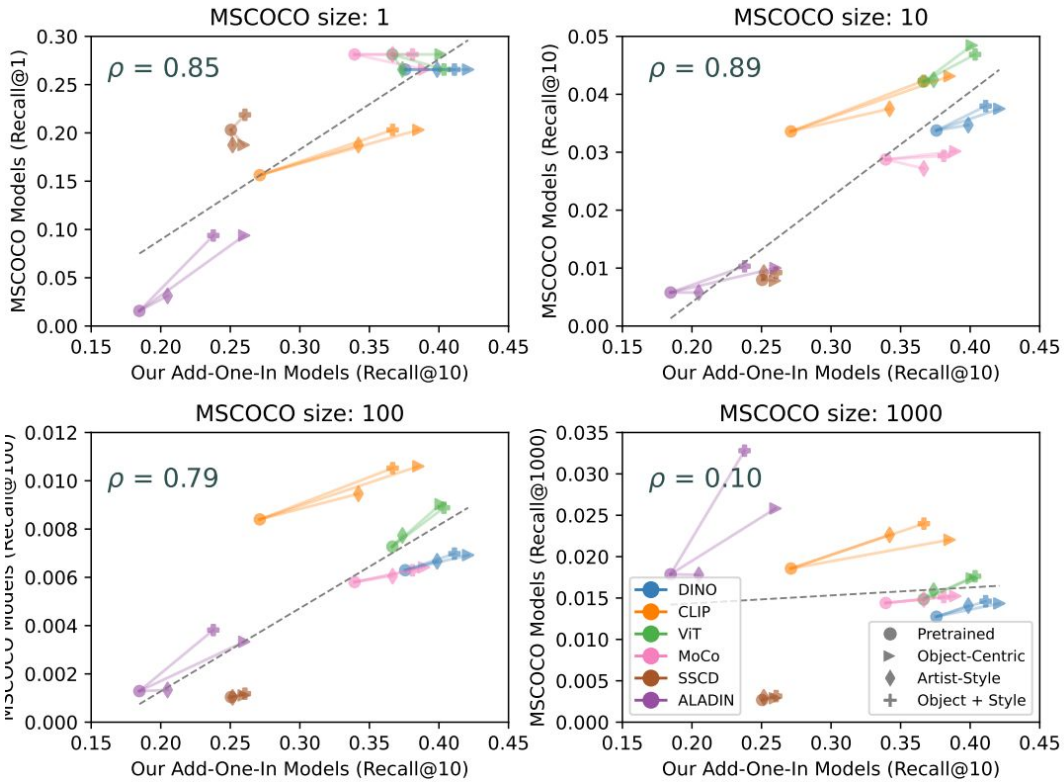


Experiments



Experiments

Single image finetune is boring. Let's do multiple!



Extra

Chat-GPT prompt:

Provide 25 diverse image captions depicting images containing **category**, where the word "**category**" is in each caption as a subject. Each caption should be applicable to depict images containing any kinds of **category** in general, without explicitly mentioning any specific **category**. Each caption should be suitable to generate realistic images using a large-scale text-to-image generative model.

Dataset examples (and more)
Prompts examples