Synthetic Data: The New Frontier

A lifelong learning guide into harnessing generative models powers

Diganta Misra, DLCT, 12th April, 2024.

ML Collective

Quick Eye Test 👀

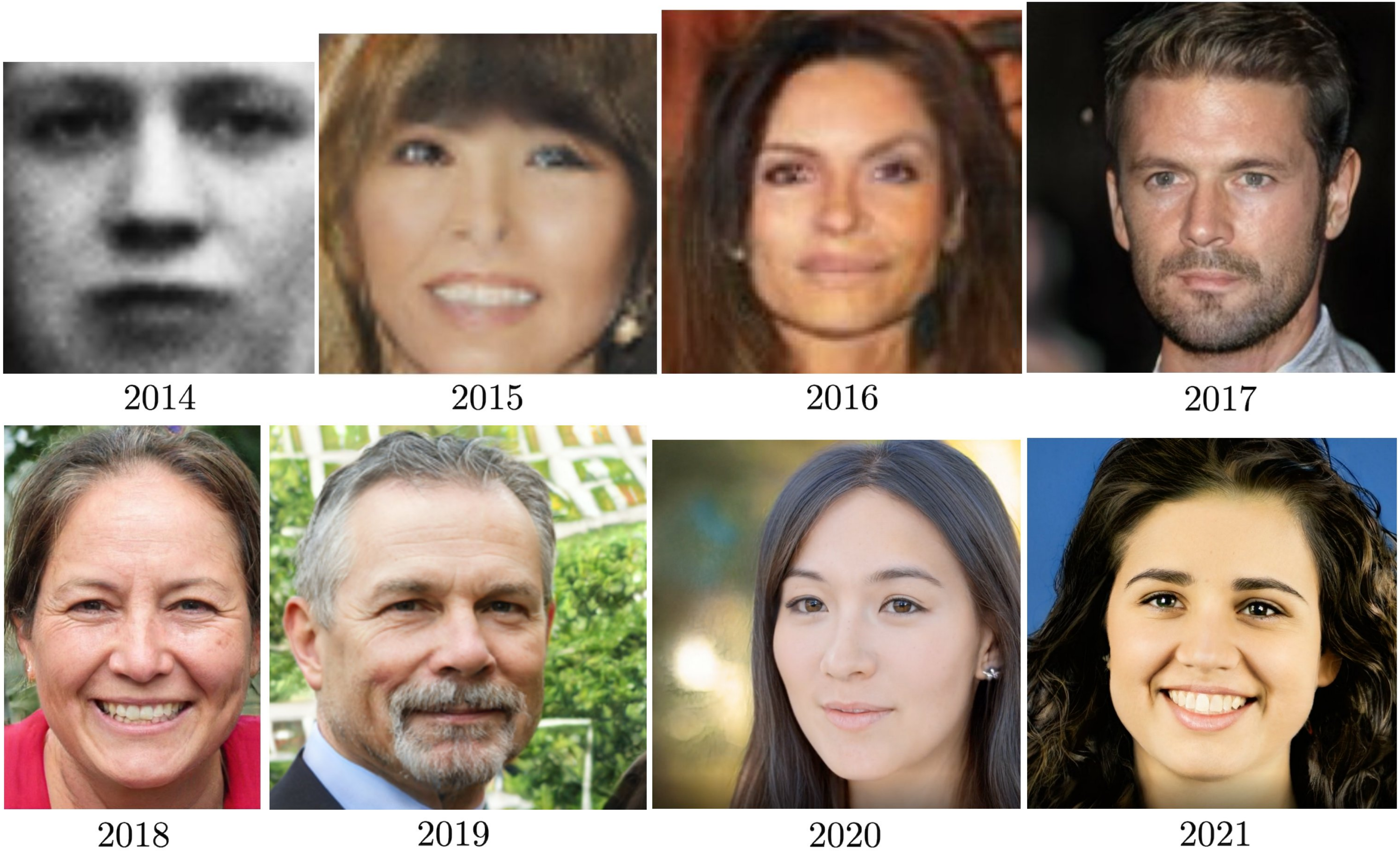
Charlie

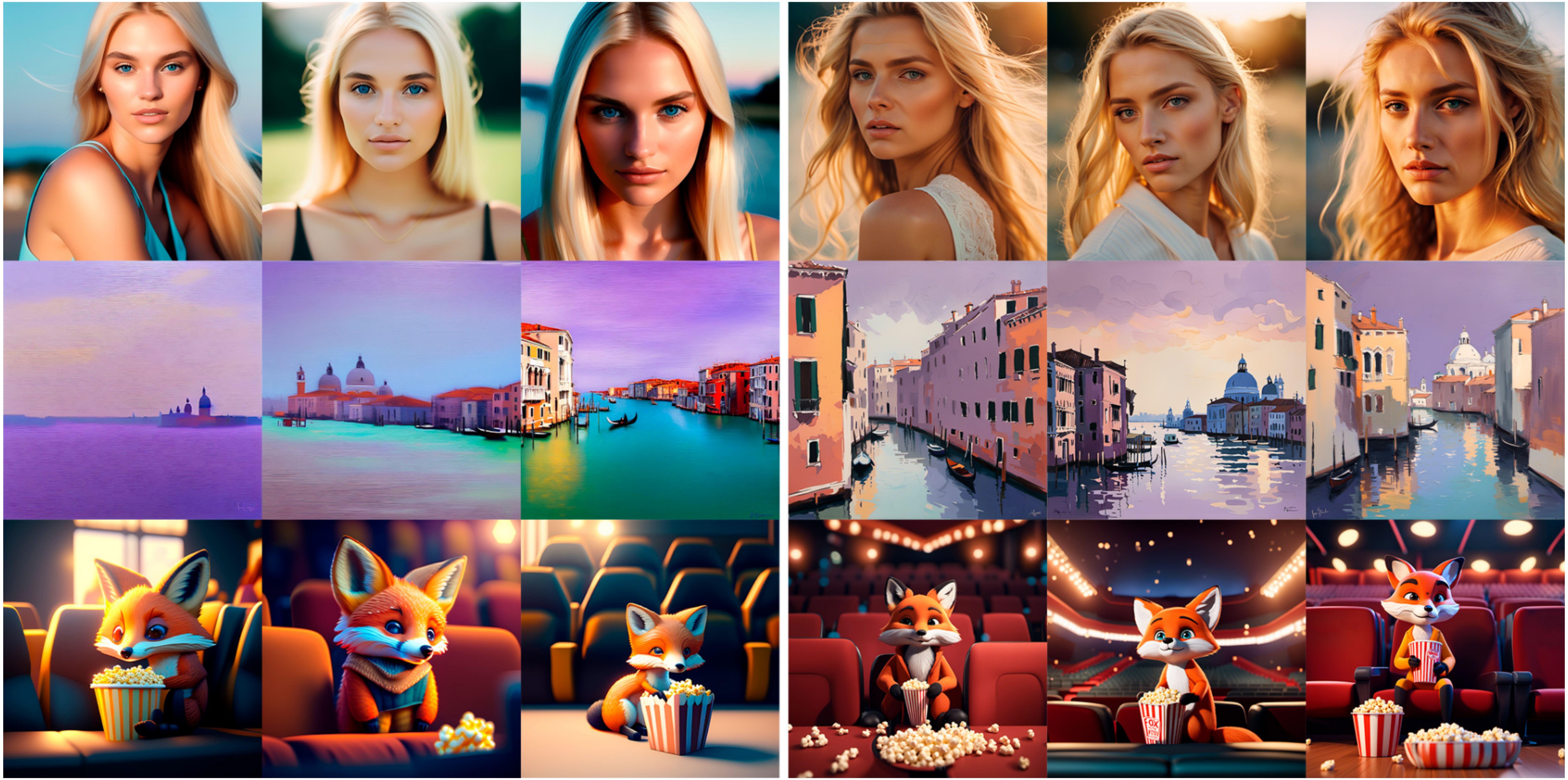
Foxtrot

Daisy

Data is wealth; generative data is its exponential growth, expanding the horizons of understanding.

ChatGPT, circa 2024


2014
2015
2016
2017
2018
2019
2020
2021



Companies are betting big

Google

s.

runway

Meta

From Supervised Training to Generative Supervised Training

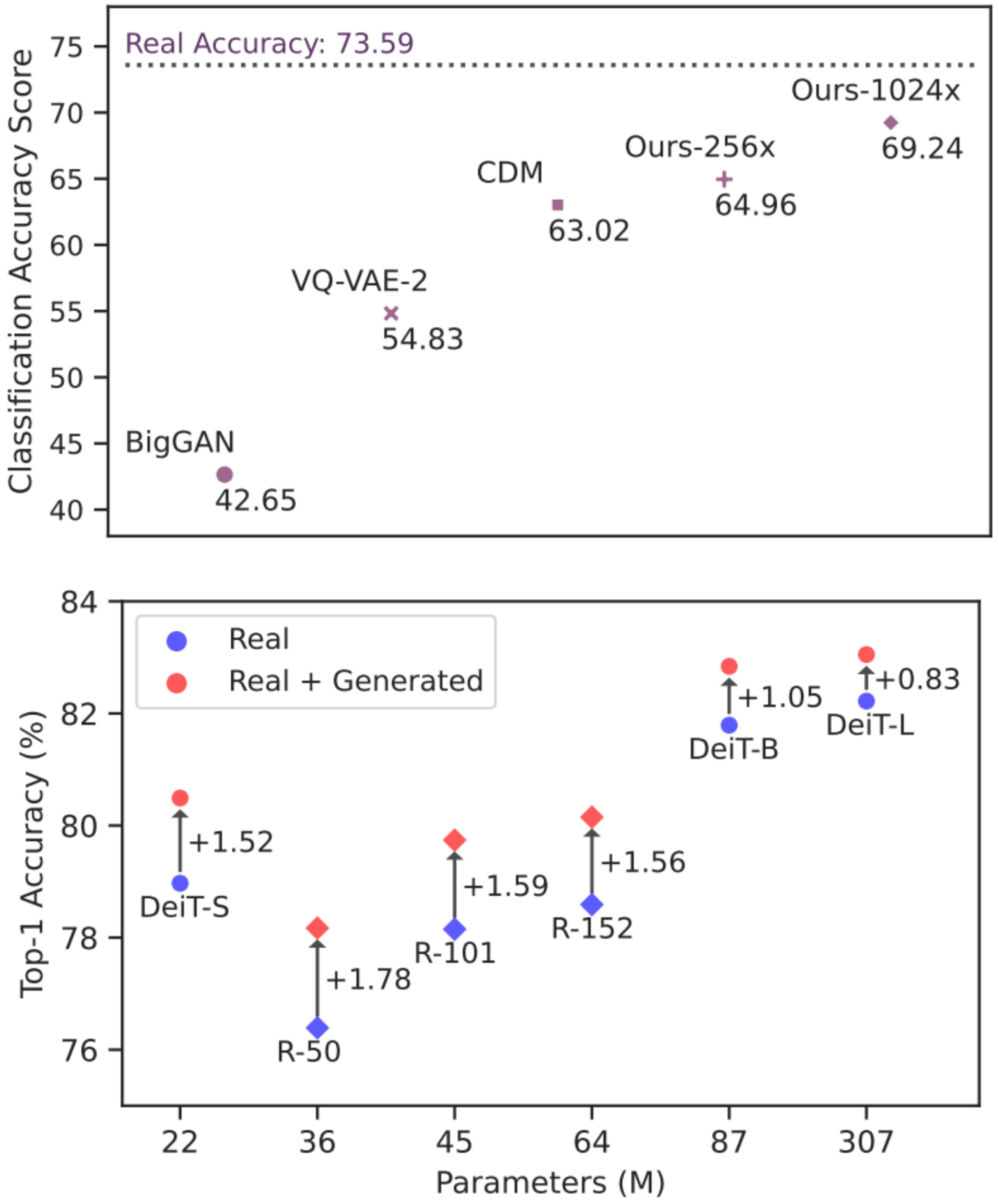
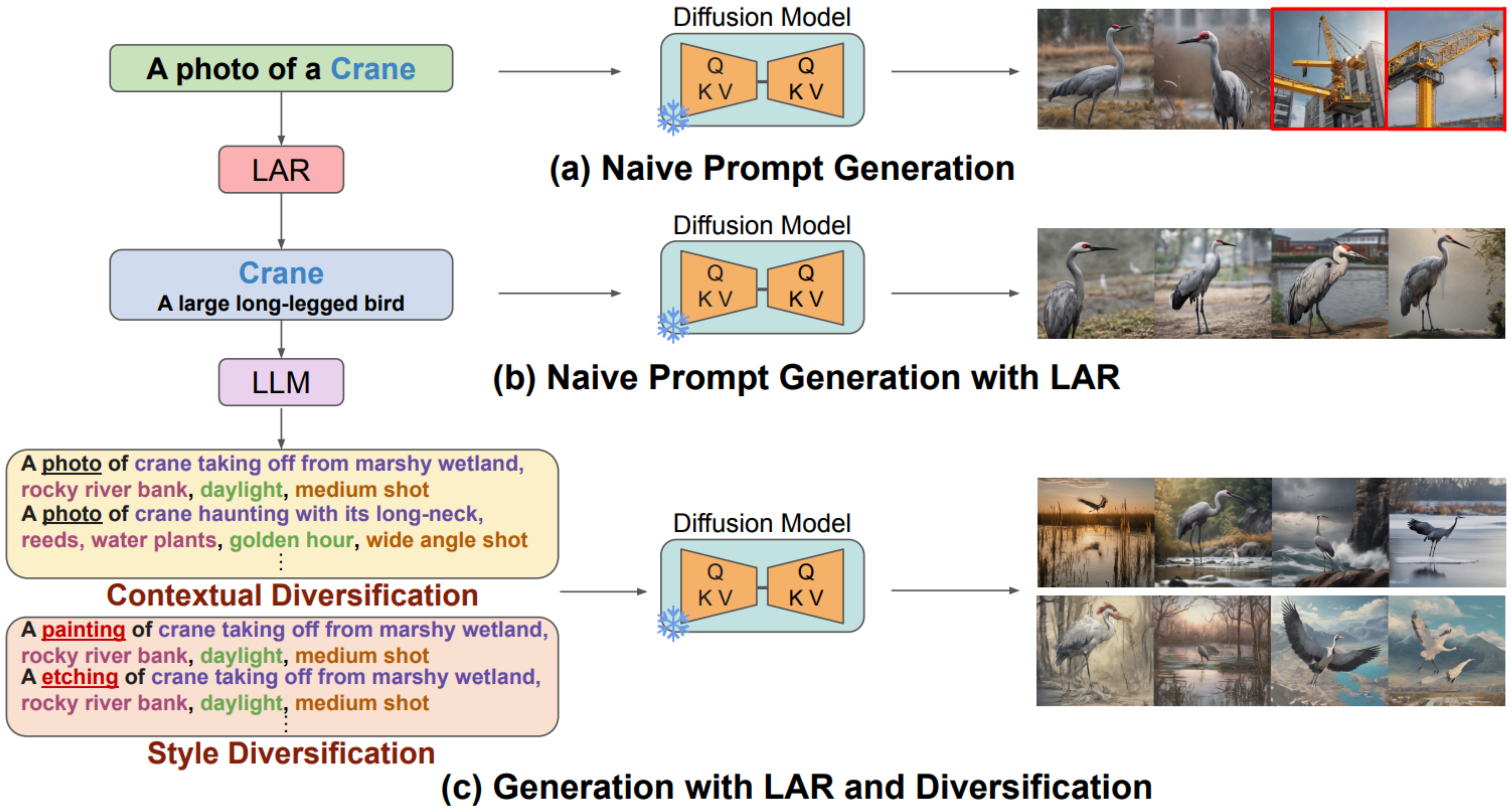


Figure 1. Top: Classification Accuracy Scores [45] show that models trained on generated data are approaching those trained on real data. Bottom: Augmenting real training data with generated images from our ImageNet model boosts classification accuracy for ResNet and Transformer models.



Azizi, Shekoofeh, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. "Synthetic data from diffusion models improves imagenet classification." arXiv preprint arXiv:2304.08466 (2023).

Yu, Zhuoran, Chenchen Zhu, Sean Culatana, Raghuraman Krishnamoorthi, Fanyi Xiao, and Yong Jae Lee. "Diversify, Don't Fine-Tune: Scaling Up Visual Recognition Training with Synthetic Images." *arXiv preprint arXiv:2312.02253* (2023).

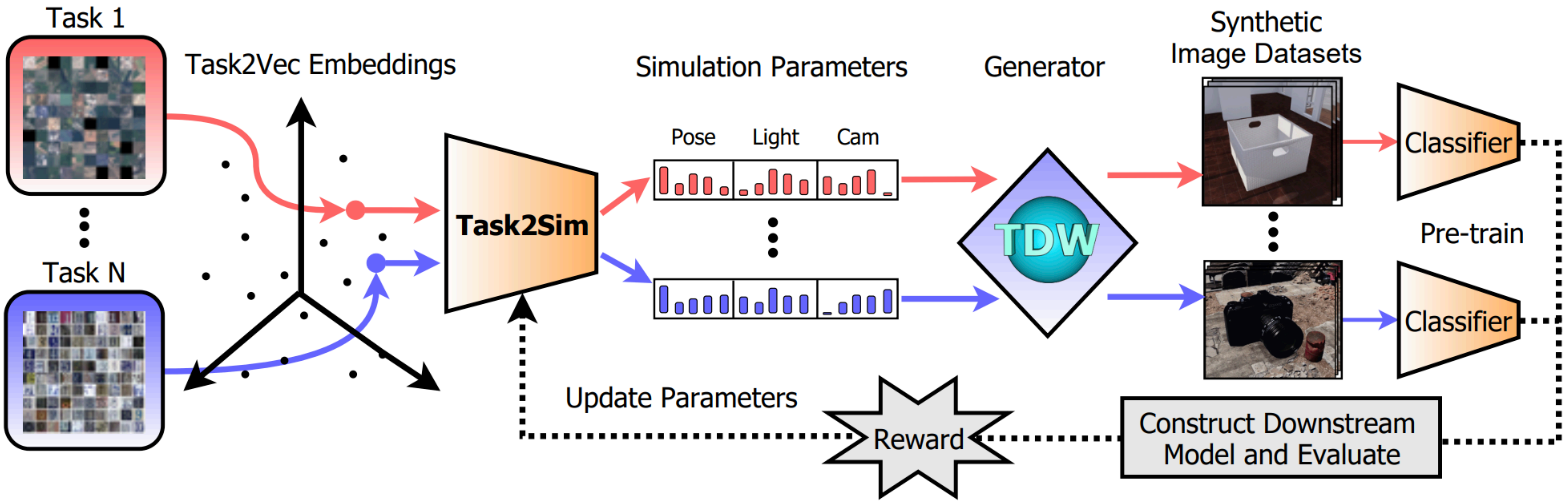


Figure 2. **Illustration of our proposed approach**. Given a batch of tasks represented by Task2Vec representations, our approach (Task2Sim) aims to map these representations to optimal simulation parameters for generating a dataset of synthetic images. The down-stream classifier's accuracy for the set of tasks is then used as a reward to update Task2Sim's parameters. Once trained, Task2Sim can be used not only for "seen" tasks but also can be used in one-shot to generate simulation parameters for novel "unseen" tasks.

Mishra, Samarth, Rameswar Panda, Cheng Perng Phoo, Chun-Fu Richard Chen, Leonid Karlinsky, Kate Saenko, Venkatesh Saligrama, and Rogerio S. Feris. "Task2sim: Towards effective pre-training and transfer from synthetic data." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9194-9204. 2022.

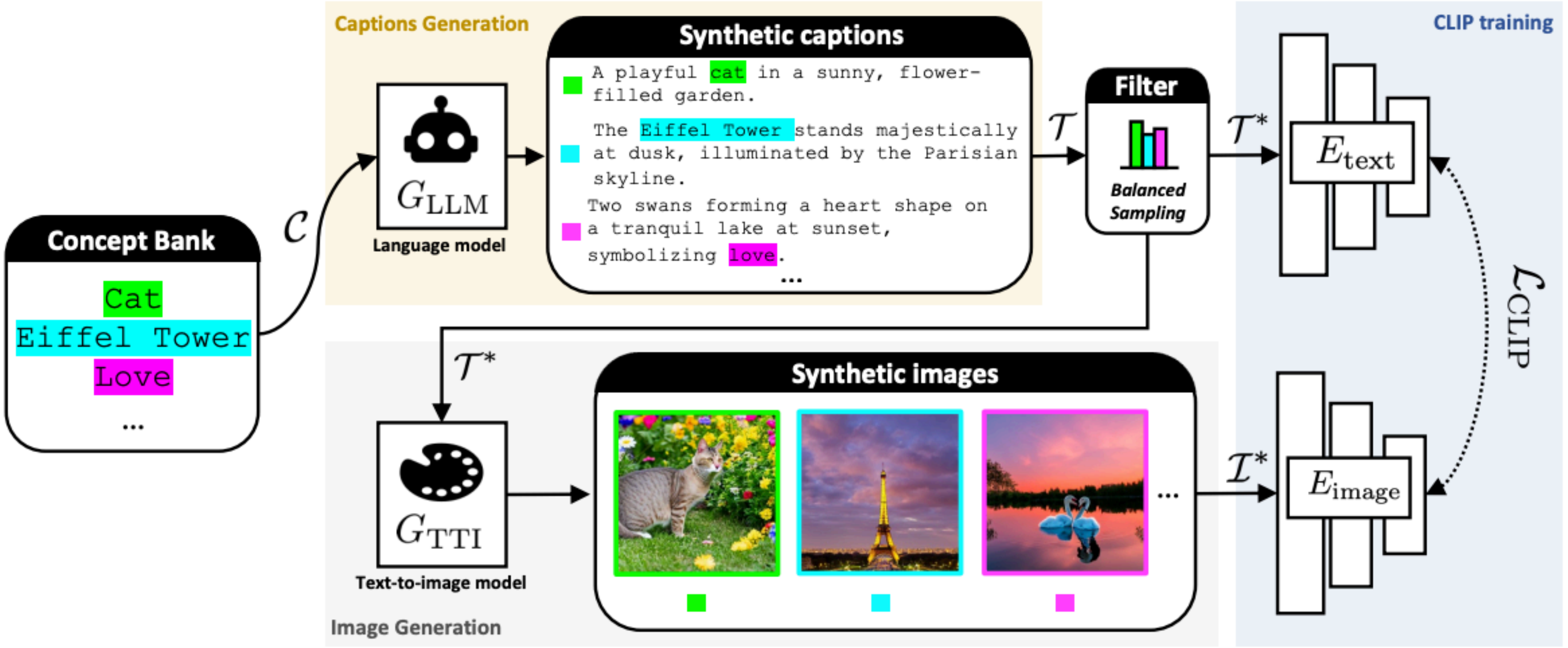


Figure 2: **Pipeline Overview.** From a set of concepts $\mathcal{C}$ (left), we obtain a set of synthetic captions $\mathcal{T}$ with an LLM, further refined to $\mathcal{T}^*$ by a filtering operation which subsamples $\mathcal{T}$ using balanced sampling (top). The generated captions are then used to prompt a text-to-image model, obtaining synthetic images aligned with the prompt (bottom). Finally, we train CLIP encoders on the generated synthetic text-image pairs. (right)

Hammoud, Hasan Abed Al Kader, Hani Itani, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. "SynthCLIP: Are We Ready for a Fully Synthetic CLIP Training?." *arXiv preprint arXiv:2402.01832* (2024).

Exploring the application of synthetic audio in training keyword spotters

Andrew Werchniak[1], *Roberto Barra Chicote*[1], *Yuriy Mischenko*[1], *Jasha Droppo*[1], *Jeff Condal*[1], *Peng Liu*[1], *Anish Shah*[1]

[1]Alexa Speech, Amazon.com

{wercha, rchicote, yuriym, drojasha, jccondal, liupng, anishsh}@amazon.com

Abstract

The study of keyword spotting, a subfield within the broader field of speech recognition that centers around identifying individual keywords in speech audio, has gained particular importance in recent years with the rise of personal voice assistants such as Alexa. As voice assistants aim to rapidly expand to support new languages, keywords, and use cases, stakeholders face the issue of limited training data for these unseen scenarios. This paper details some initial exploration into the application of Text-To-Speech (TTS) audio as a "helper" tool for training keyword spotters in these low-resource scenarios. In the experiments studied in this paper, the careful mixing of TTS audio with human speech audio during training led to a reduction of over 11% in the detection-error-tradeoff (DET) area under the curve (AUC) metric.

Index Terms: keyword spotting, speech recognition, data augmentation, speech synthesis

1. Introduction

Over the past few years, voice assistants such as Amazon's Alexa, Google Assistant, and Apple's Siri have risen rapidly in popularity, to the point that they have become a staple of everyday life for many people across the globe. Alexa, in particular, now has tens of millions of users who interact with their

including the data preparation, model training, and model evaluation; Section 4 details the experimental results; and Section 5 summarizes the conclusions and future work to build on the results.

2. Related Work

Some previous research has been dedicated to the application of synthetic audio in training automatic speech recognition (ASR) systems. Large vocabulary ASR models of architectures varying from Gaussian Mixture Models (GMM)/Hidden Markov Models (HMM) [5] to Convolutional Neural Network(CNN)/Connectionist Temporal Classification (CTC) models [6] to more modern attention-based acoustic-to-word models [7, 8] have all been shown to benefit from the addition of TTS data at varying levels and stages. However, it is worth noting that there may be limits to these benefits, as it has been shown that bispectral analysis can still differentiate with confidence between audio generated with state-of-the-art TTS systems and human audio[9], indicating that a mismatch may still exist between synthetic training audio and organic evaluation audio.

Regardless, the application of synthetic data in training low-resource keyword spotter systems has shown promise in recent experiments. Specifically, it was demonstrated that by utilizing a pre-trained speech-embedding model with approximately 400K parameters and weights initialized using human

Werchniak, Andrew, Roberto Barra Chicote, Yuriy Mishchenko, Jasha Droppo, Jeff Condal, Peng Liu, and Anish Shah. "Exploring the application of synthetic audio in training keyword spotters." In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7993-7996. IEEE, 2021.

Published as a conference paper at ICLR 2024

Pre-training with Synthetic Data Helps Offline Reinforcement Learning

Zecheng Wang[1*‡] **Che Wang**[2,4*†] **Zixuan Dong**[3,4*] **Keith Ross**[1]

[1] New York University Abu Dhabi [2] New York University Shanghai
[3] SFSC of AI and DL, NYU Shanghai [4] New York University

Abstract

Recently, it has been shown that for offline deep reinforcement learning (DRL), pre-training Decision Transformer with a large language corpus can improve downstream performance (Reid et al., 2022). A natural question to ask is whether this performance gain can only be achieved with language pre-training, or can be achieved with simpler pre-training schemes which do not involve language. In this paper, we first show that language is not essential for improved performance, and indeed pre-training with synthetic IID data for a small number of updates can match the performance gains from pre-training with a large language corpus; moreover, pre-training with data generated by a one-step Markov chain can further improve the performance. Inspired by these experimental results, we then consider pre-training Conservative Q-Learning (CQL), a popular offline DRL algorithm, which is Q-learning-based and typically employs a Multi-Layer Perceptron (MLP) backbone. Surprisingly, pre-training with simple synthetic data for a small number of updates can also improve CQL, providing consistent performance improvement on D4RL Gym locomotion datasets. The results of this paper not only illustrate the importance of pre-training for offline DRL but also show that the pre-training data can be synthetic and generated with remarkably simple mechanisms.

Wang, Zecheng, Che Wang, Zixuan Dong, and Keith Ross. "Pre-training with Synthetic Data Helps Offline Reinforcement Learning." *arXiv preprint arXiv:2310.00771* (2023).

Does Generated Data Scale?

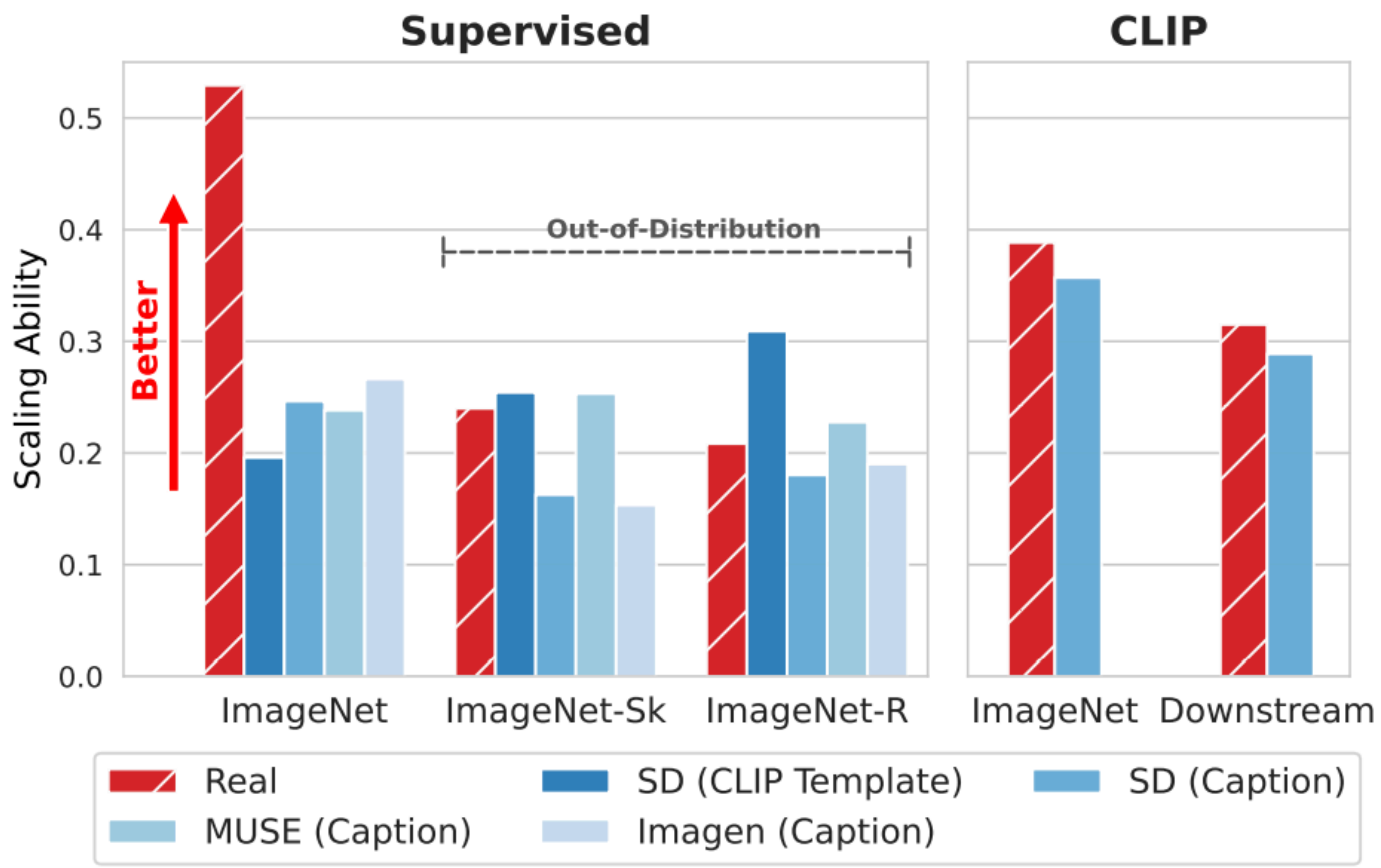


Figure 1. Scaling ability (*i.e.*, the slope of the power law curve between loss and dataset size fitted in the log space, see Eq. 2) comparison between real and synthetic images on supervised classifier and CLIP training. Red bars represent real images and blue bars represent synthetic images generated with different text-to-image models. Supervised models are trained on real or synthetic ImageNet, and text in parentheses is the text prompt used to generate the images (details in Section 3.1). ImageNet-Sketch and ImageNet-R are out-of-distribution tests. CLIP models are trained on LAION-400M with real or synthetic images. We see that: (1) scaling ability of synthetic data is *slightly worse* than that of real data for CLIP training; (2) robustness on ImageNet-Sketch and ImageNet-R datasets can be *better* when training on synthetic data.

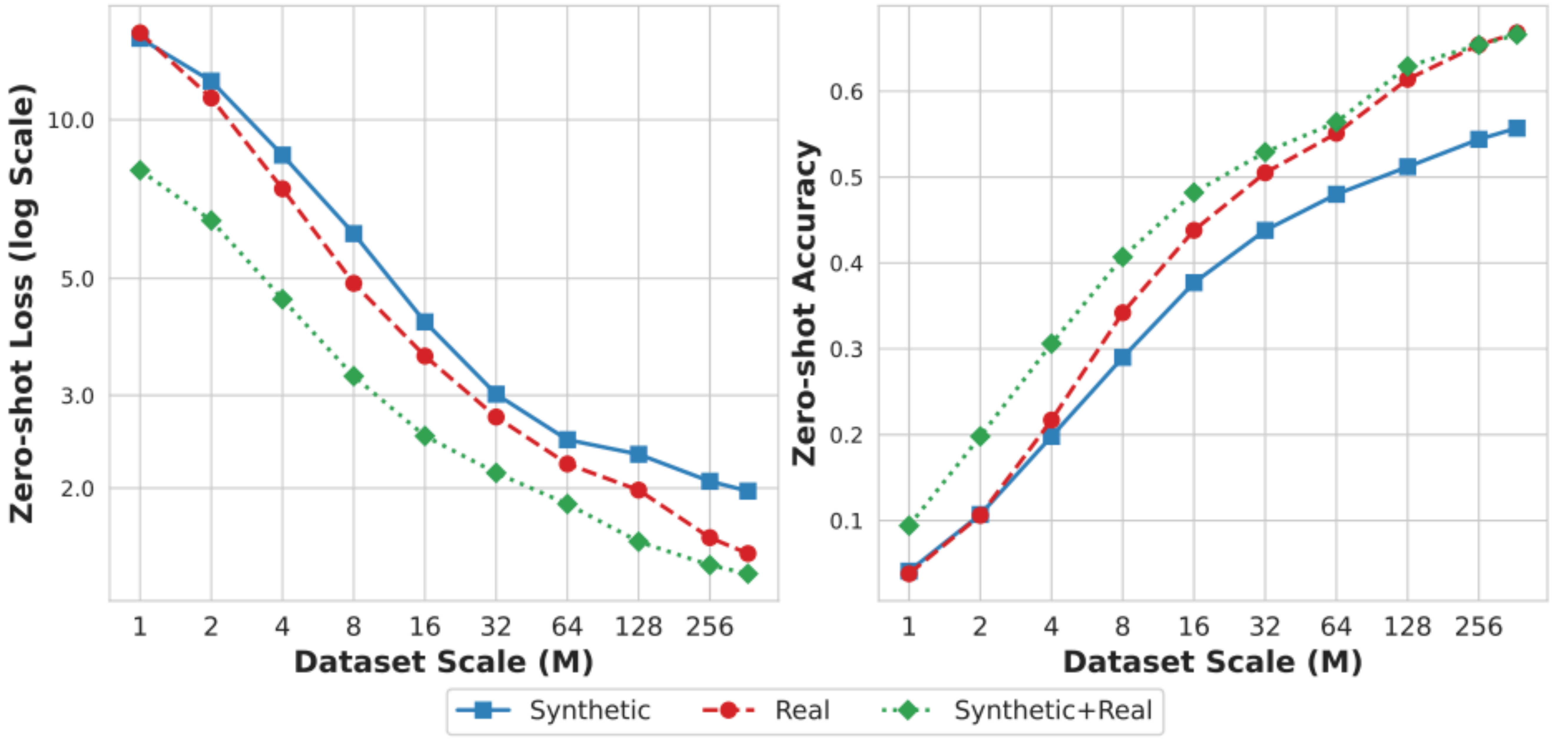


Figure 9. Scaling behavior for CLIP models trained on LAION-400M subsets of different scales. Models are trained with synthetic, real, or a combination of synthetic and real images, and are evaluated with ImageNet zero-shot accuracy. Dataset scale here refers to the number of captions.

Fan, Lijie, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. "Scaling laws of synthetic images for model training... for now." *arXiv preprint arXiv:2312.04567* (2023).

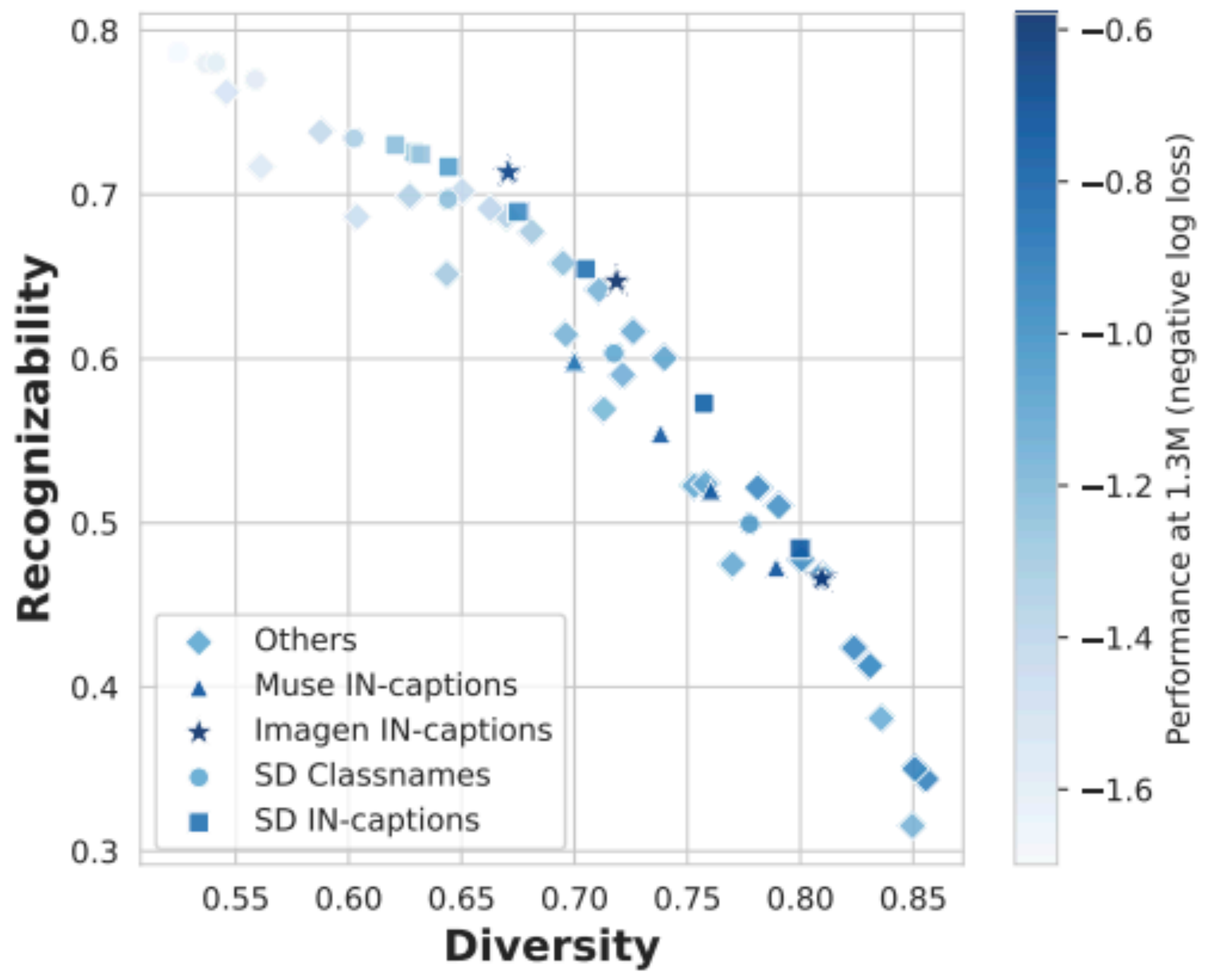


Figure 2. Recognizability vs. diversity plot for various synthetic image generation configurations (as in Section 4.2), colored by the performance at 1.3M on ImageNet validation set (measured by negative log loss). Deeper color stands for smaller loss and better performance.

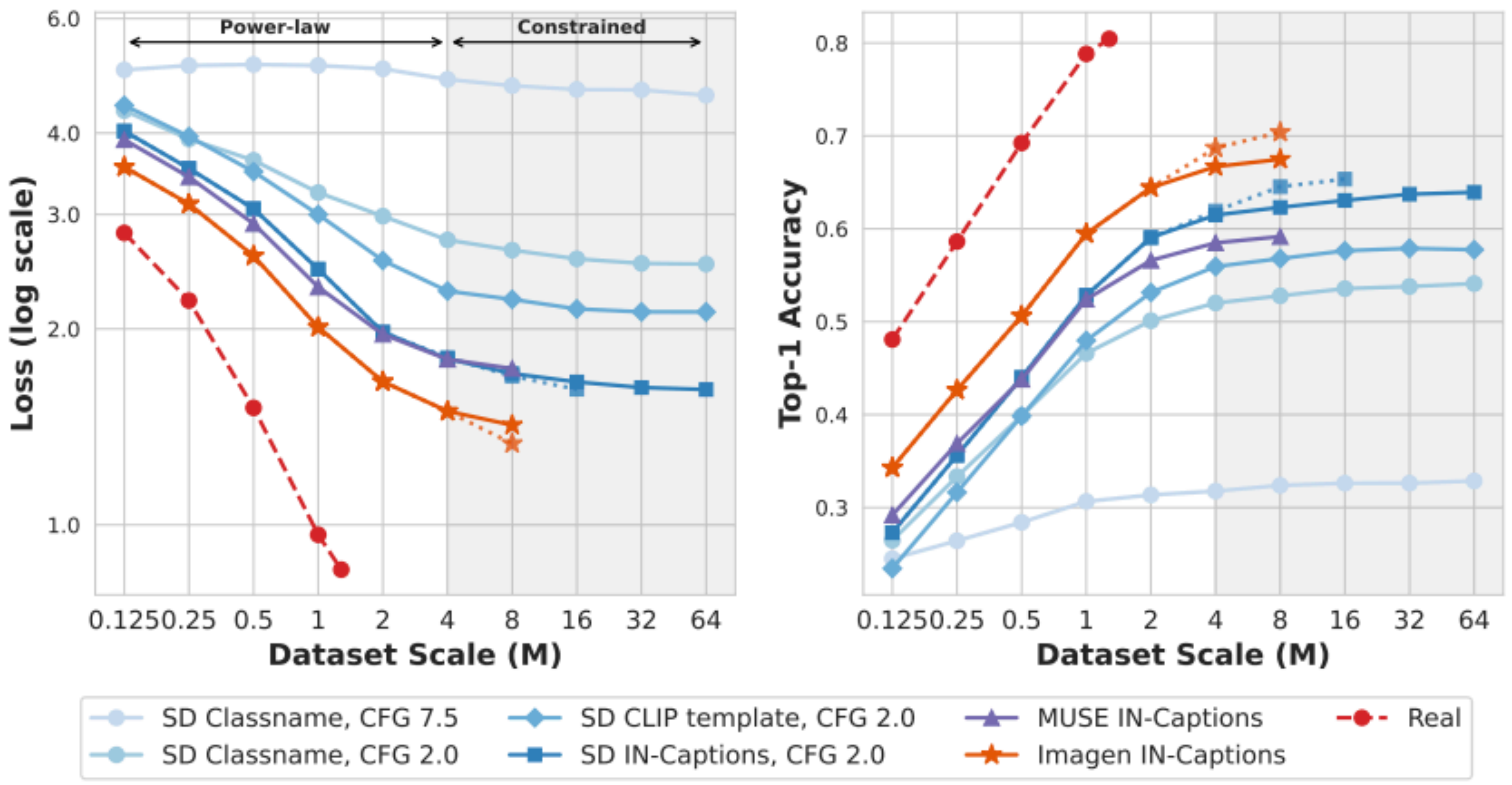


Figure 3. Scaling on ImageNet validation set for various configurations as in Section 4.3. Loss and data scale follows the power-law (as in Equation 2) with varied k when data is less than 4M. By tuning the CFG scale, text prompts and text-to-image models, the scaling behavior for synthetic images can be significantly improved (from light blue to orange). Red dashed line is for real images. Orange and blue dotted lines are ViT-L backbones, extending the power-law to 8M.

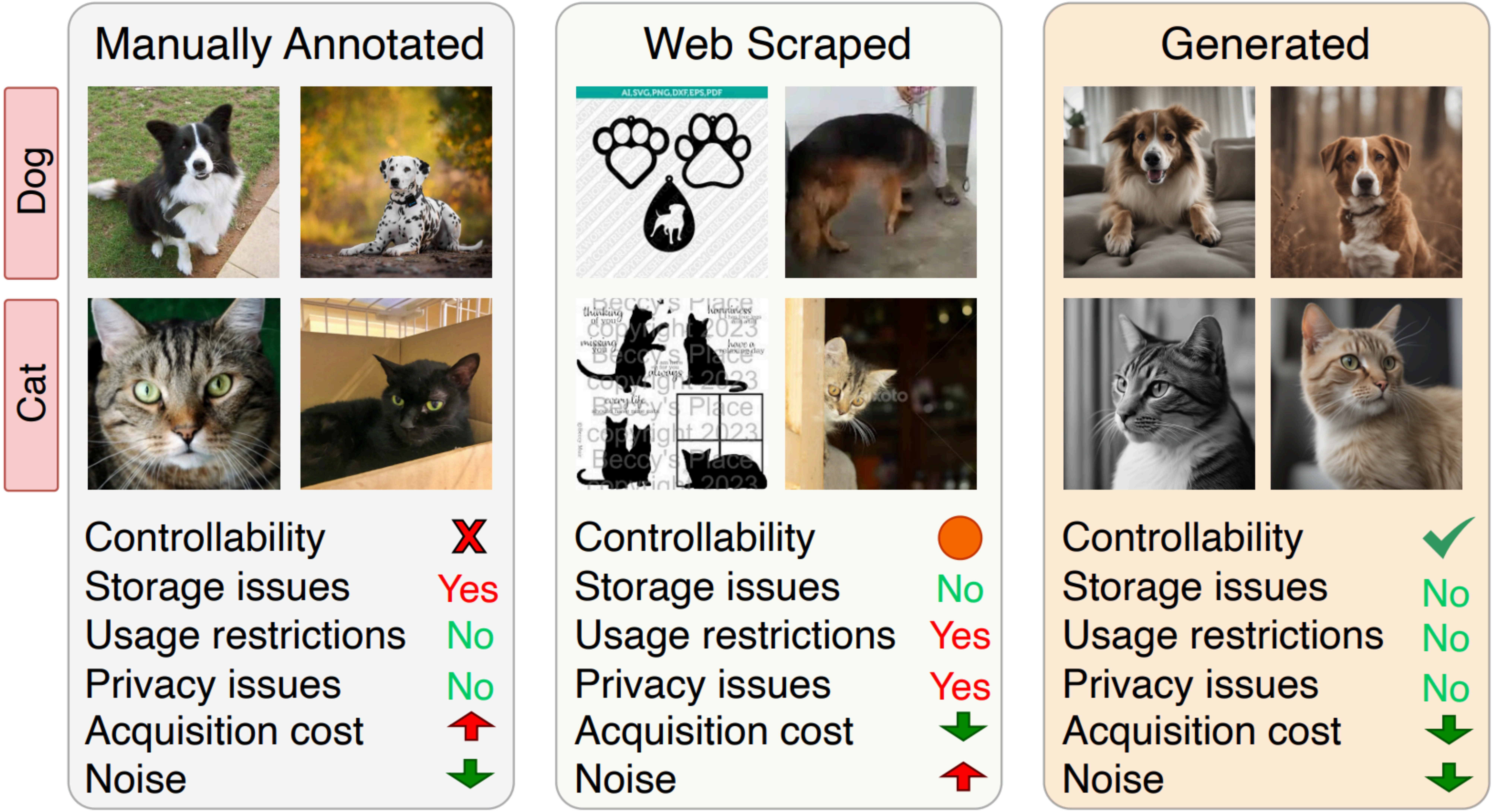
Fan, Lijie, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. "Scaling laws of synthetic images for model training... for now." *arXiv preprint arXiv:2312.04567* (2023).

Just Say the Name: Online Continual Learning with Category Names Only via Data Generation

Minhyuk Seo, Diganta Misra, Seongwon Cho, Minjae Lee, Jonghyun Choi

Seo, Minhyuk, Diganta Misra, Seongwon Cho, Minjae Lee, and Jonghyun Choi. "Just Say the Name: Online Continual Learning with Category Names Only via Data Generation." *arXiv preprint arXiv:2403.10853* (2024).

Background



Manually Annotated
Web Scraped
Generated
Dog
Cat
Controllability
Storage issues
Yes
Usage restrictions
No
Privacy issues
No
Acquisition cost
Noise
Controllability
Storage issues
No
Usage restrictions
Yes
Privacy issues
Yes
Acquisition cost
Noise
Controllability
Storage issues
No
Usage restrictions
No
Privacy issues
No
Acquisition cost
Noise

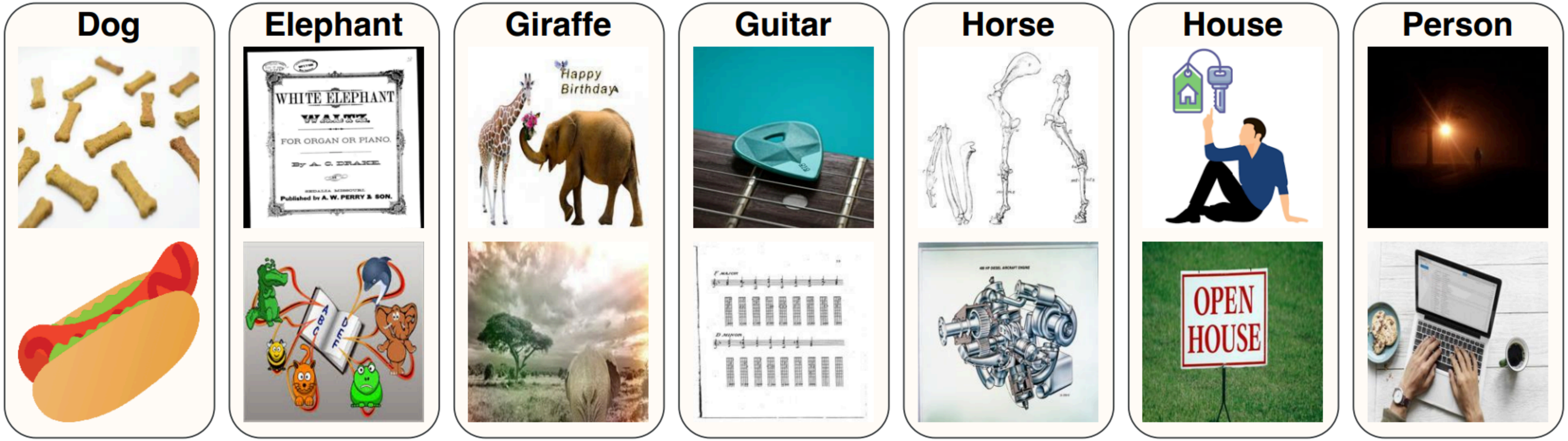
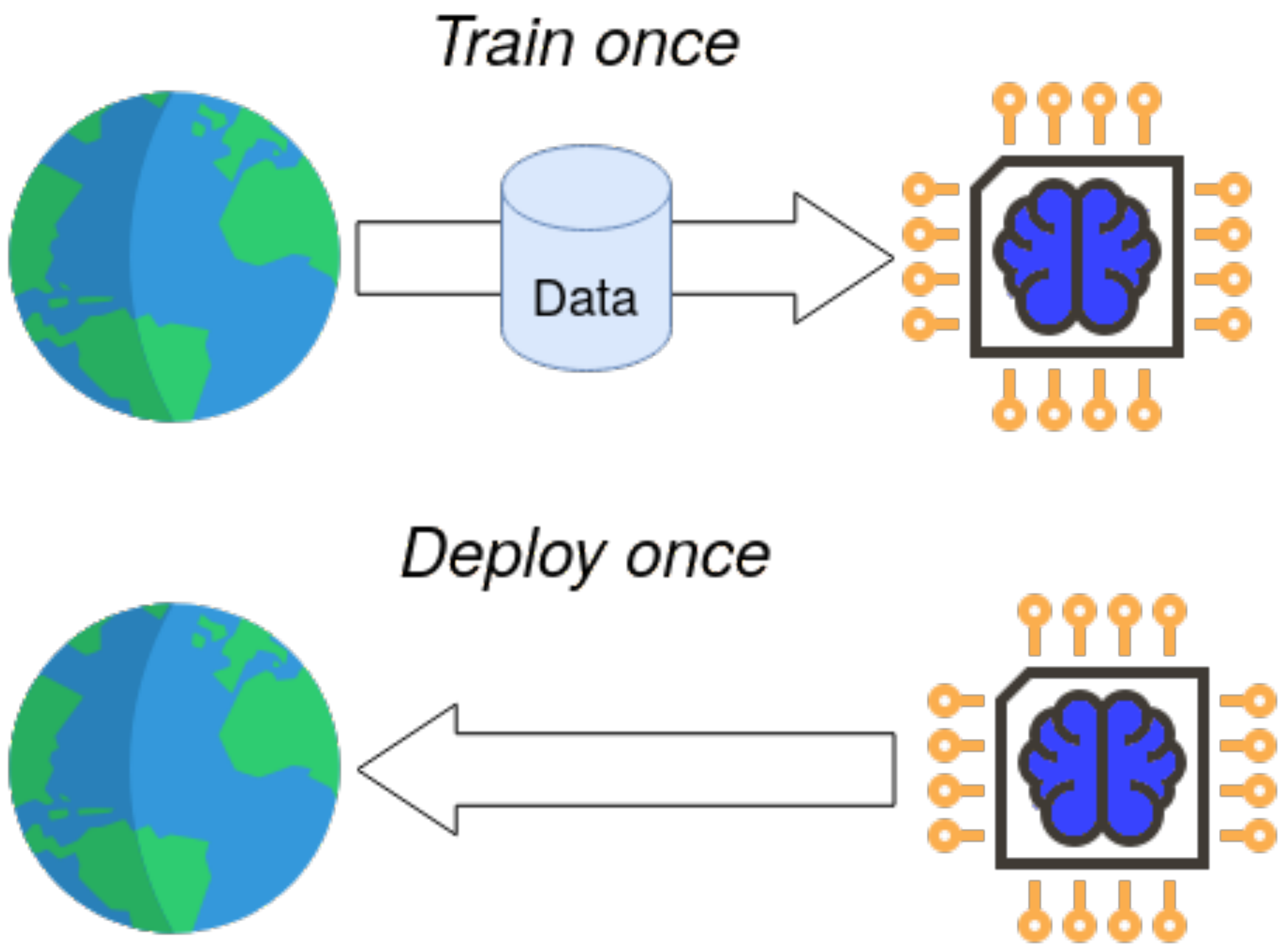
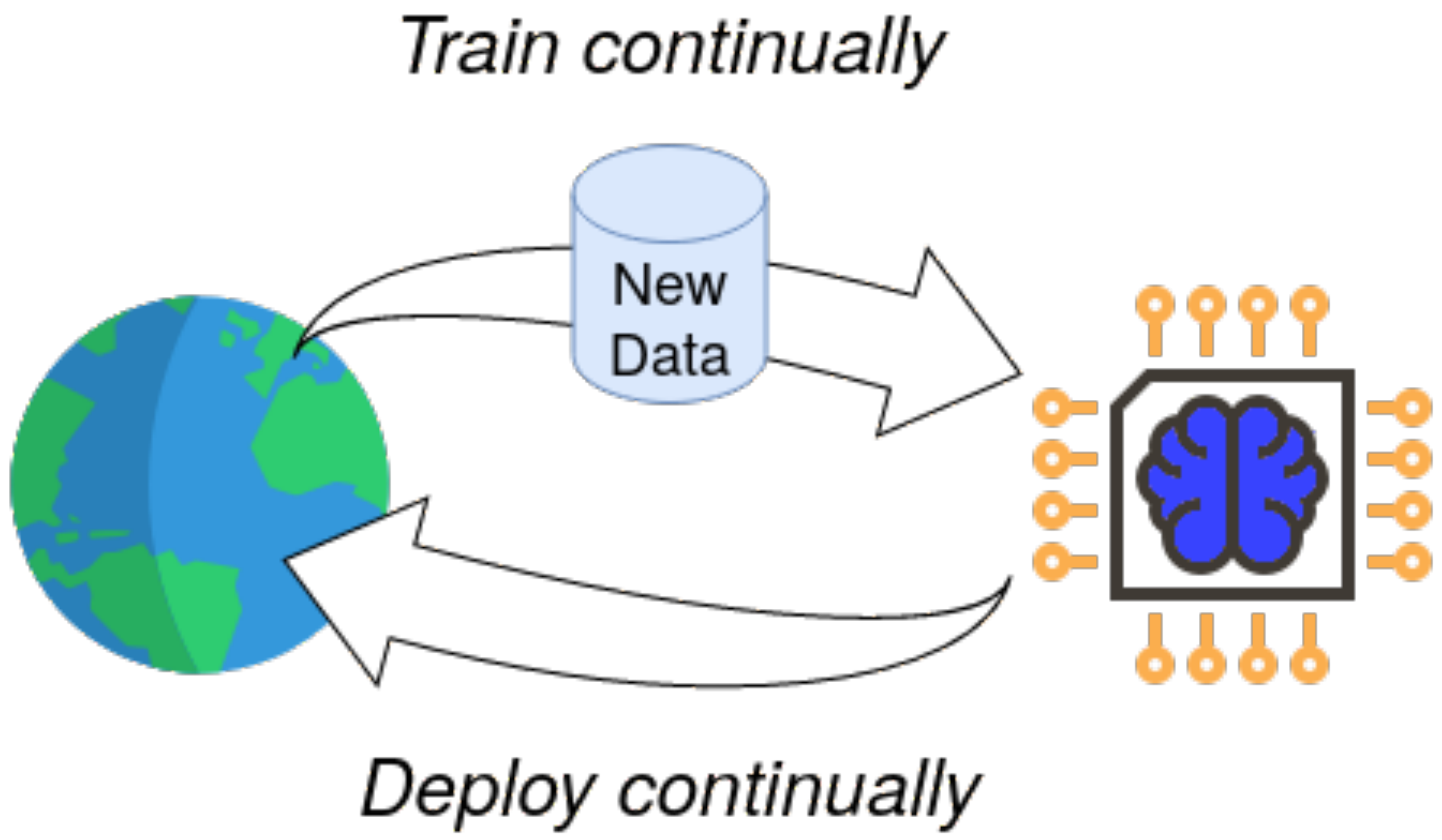


Fig. 9: Examples of noisy raw data obtained via web-scraping for the classes in the PACS [144] dataset.

Traditional ML



Continual Learning



From Categories to Classifier: Name-Only Continual Learning by Exploring the Web

Ameya Prabhu[1]* **Hasan Abed Al Kader Hammoud**[1,2]* **Ser-Nam Lim**[3] **Bernard Ghanem**[2]
Philip H.S. Torr[1] **Adel Bibi**[1]

[1]University of Oxford [2]KAUST [3]Meta AI

Abstract

Continual Learning (CL) often relies on the availability of extensive annotated datasets, an assumption that is unrealistically time-consuming and costly in practice. We explore a novel paradigm termed *name-only continual learning* where time and cost constraints prohibit manual annotation. In this scenario, learners adapt to new category shifts using only category names without the luxury of annotated training data. Our proposed solution leverages the expansive and ever-evolving internet to query and download *uncurated* webly-supervised data for image classification. We investigate the reliability of our web data and find them comparable, and in some cases superior, to manually annotated datasets. Additionally, we show that by harnessing the web, we can create support sets that surpass state-of-the-art name-only classification that create support sets using generative models or image retrieval from LAION-5B, achieving up to 25% boost in accuracy. When applied across varied continual learning contexts, our method consistently exhibits a small performance gap in comparison to models trained on manually annotated datasets. We present *EvoTrends*, a class-incremental dataset made from the web to capture real-world trends, created in just minutes. Overall, this paper underscores the potential of using uncurated webly-supervised data to mitigate the challenges associated with manual data labeling in continual learning.

Prabhu, Ameya, Hasan Abed Al Kader Hammoud, Ser-Nam Lim, Bernard Ghanem, Philip HS Torr, and Adel Bibi. "From Categories to Classifier: Name-Only Continual Learning by Exploring the Web." *arXiv preprint arXiv:2311.11293* (2023).

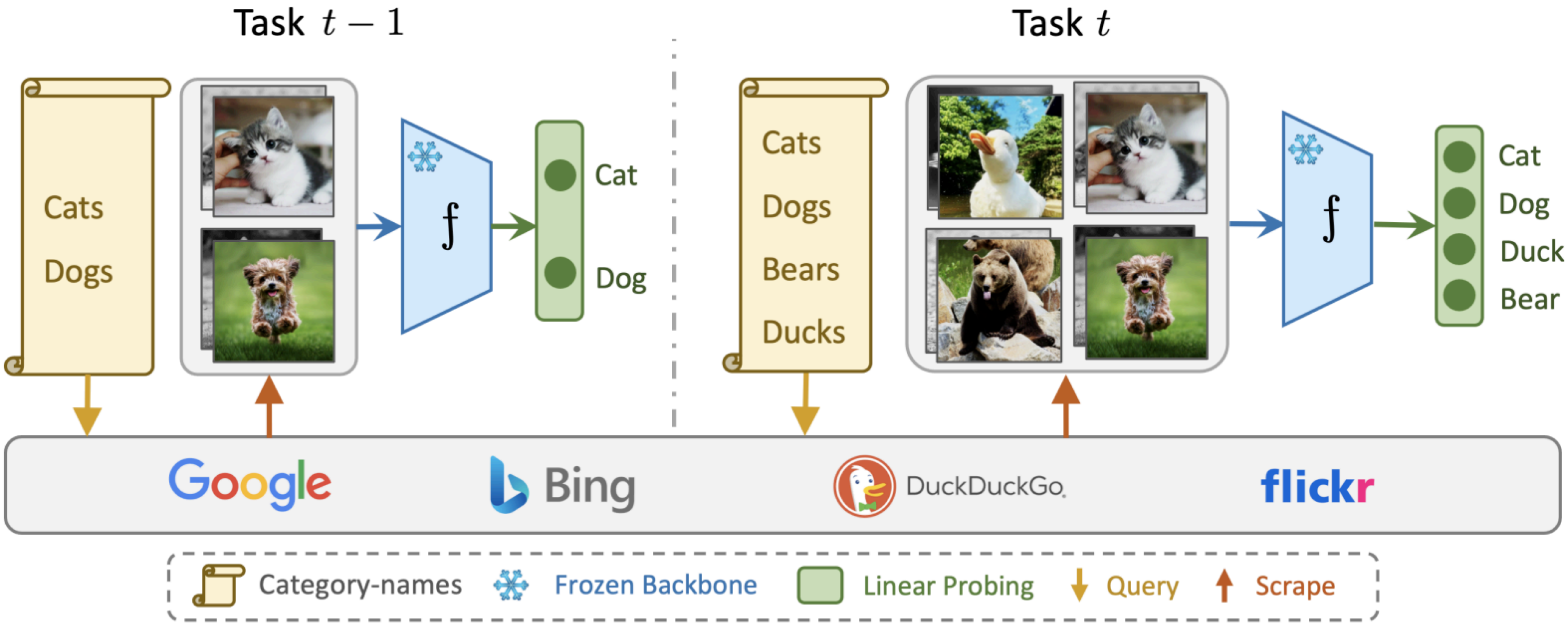
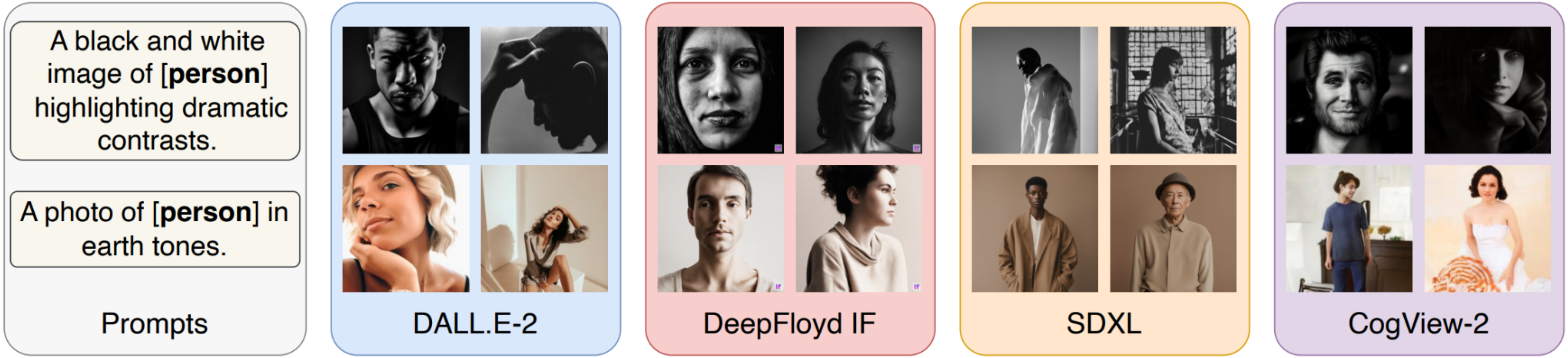


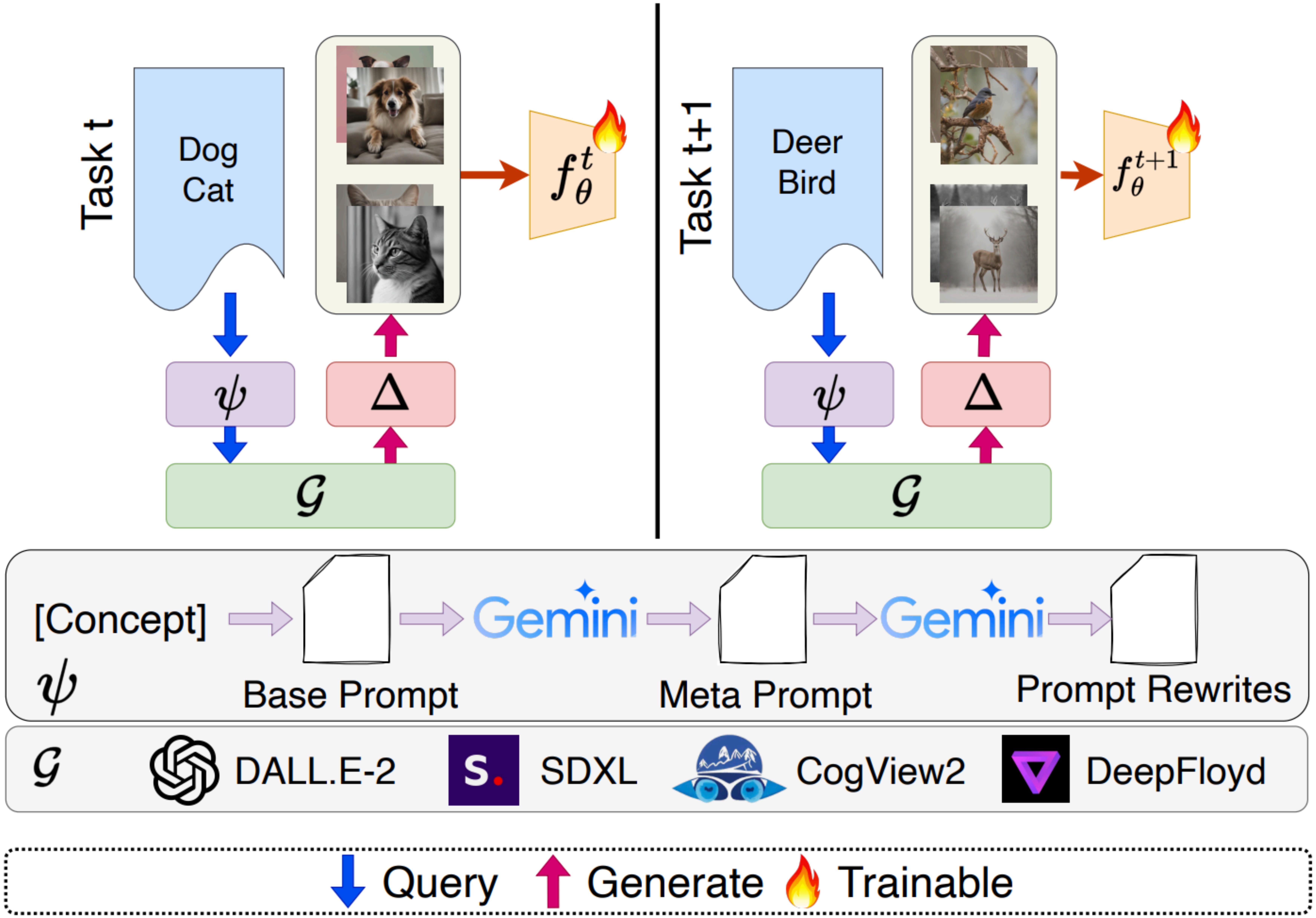
Figure 1: **Continual Name-Only Classification: Our Approach.** At each timestep t, the learner receives a list of class categories without any training samples. We start by collecting webly-supervised data through querying and downloading data from multiple search engines. We then extract features using a frozen backbone, and subsequently train a linear layer on those features. The same process is repeated for the next timestep.

Prabhu, Ameya, Hasan Abed Al Kader Hammoud, Ser-Nam Lim, Bernard Ghanem, Philip HS Torr, and Adel Bibi. "From Categories to Classifier: Name-Only Continual Learning by Exploring the Web." *arXiv preprint arXiv:2311.11293* (2023).



Previous works relied on only training from a single generator samples but what if we can couple **n** generators (specialized or generalized) and subsample from the total set?

G-NoCL


Task t
Dog
Cat
f_θ^t
ψ
Δ
$\mathcal{G}$
Task t+1
Deer
Bird
f_θ^{t+1}
ψ
Δ
$\mathcal{G}$
[Concept]
Gemini
Gemini
ψ
Base Prompt
Meta Prompt
Prompt Rewrites
$\mathcal{G}$
DALL.E-2
SDXL
CogView2
DeepFloyd
Query
Generate
Trainable

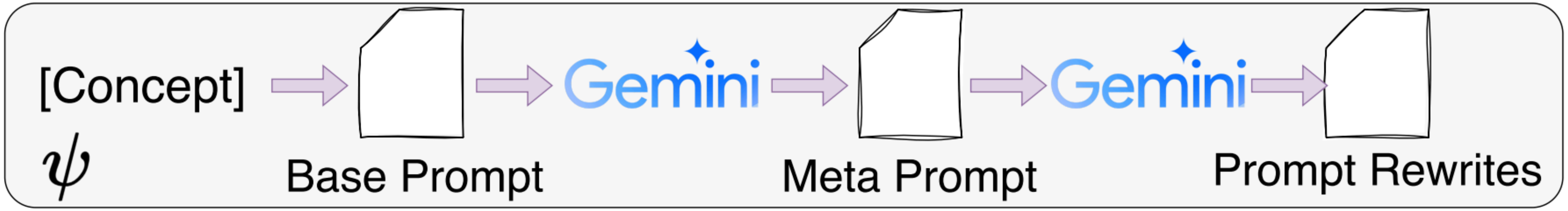
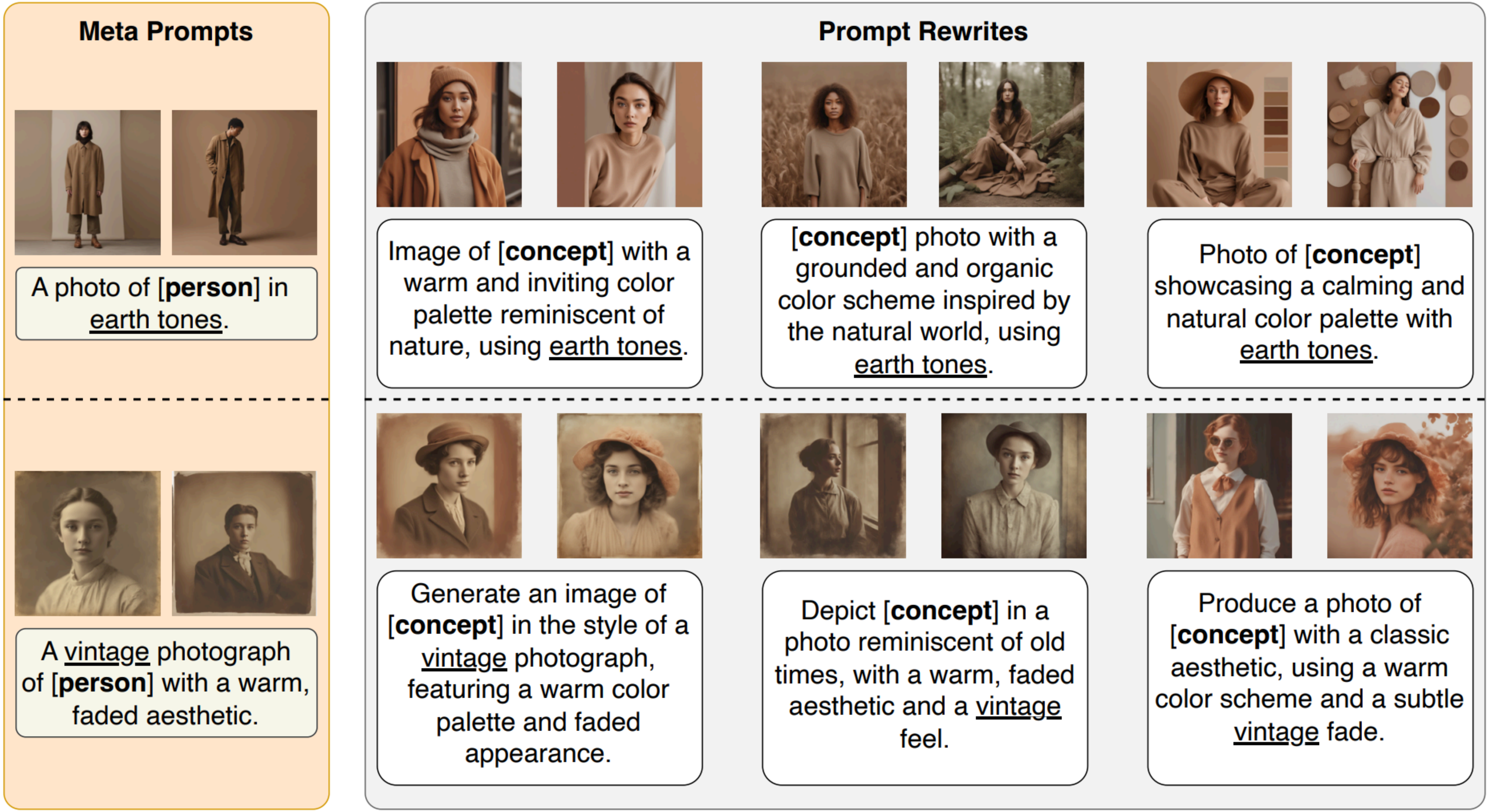


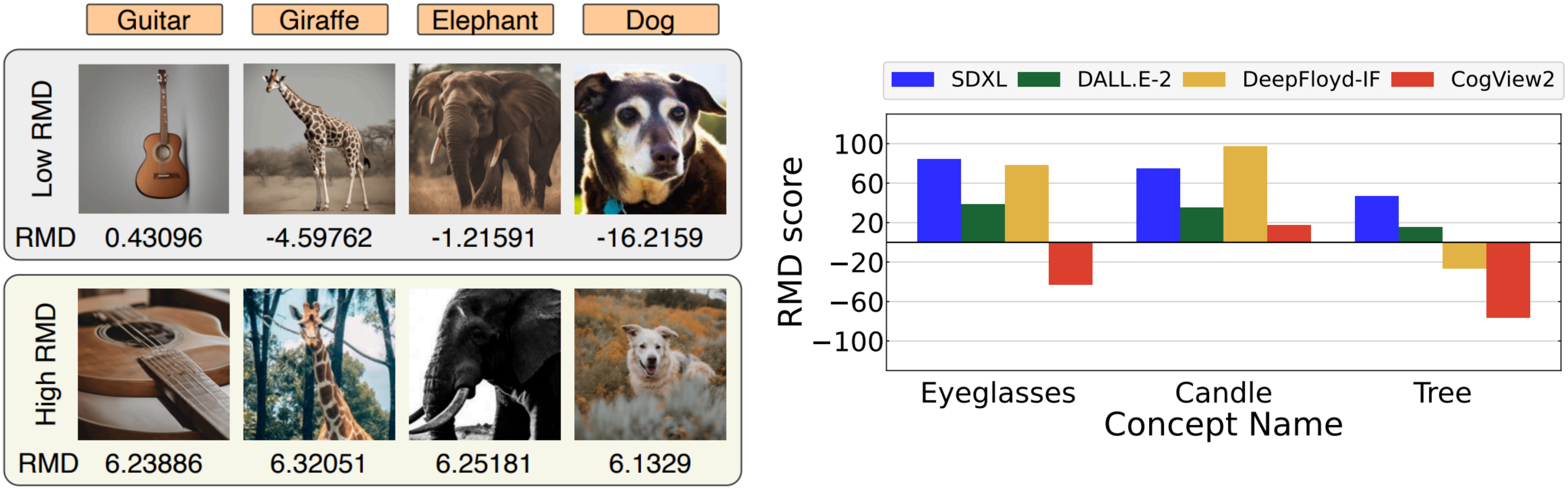
Fig. 10: Prompt Refiner Module (ψ): Given a [*concept*], ψ utilizes a pretrained frozen LLM to generate fine-grained prompt-rewrites in a two-step process.

LLM	GISTEmbed-L [115]	mxbai-embed-L[11]	Sentence-T5-B [84]	LaBSE [37]	Jina-v2 [44]
GPT-3.5 [17]	0.8552	0.8811	0.944	0.7602	0.9089
Gemini [121]	**0.8088**	**0.8544**	**0.9137**	**0.7187**	**0.8987**

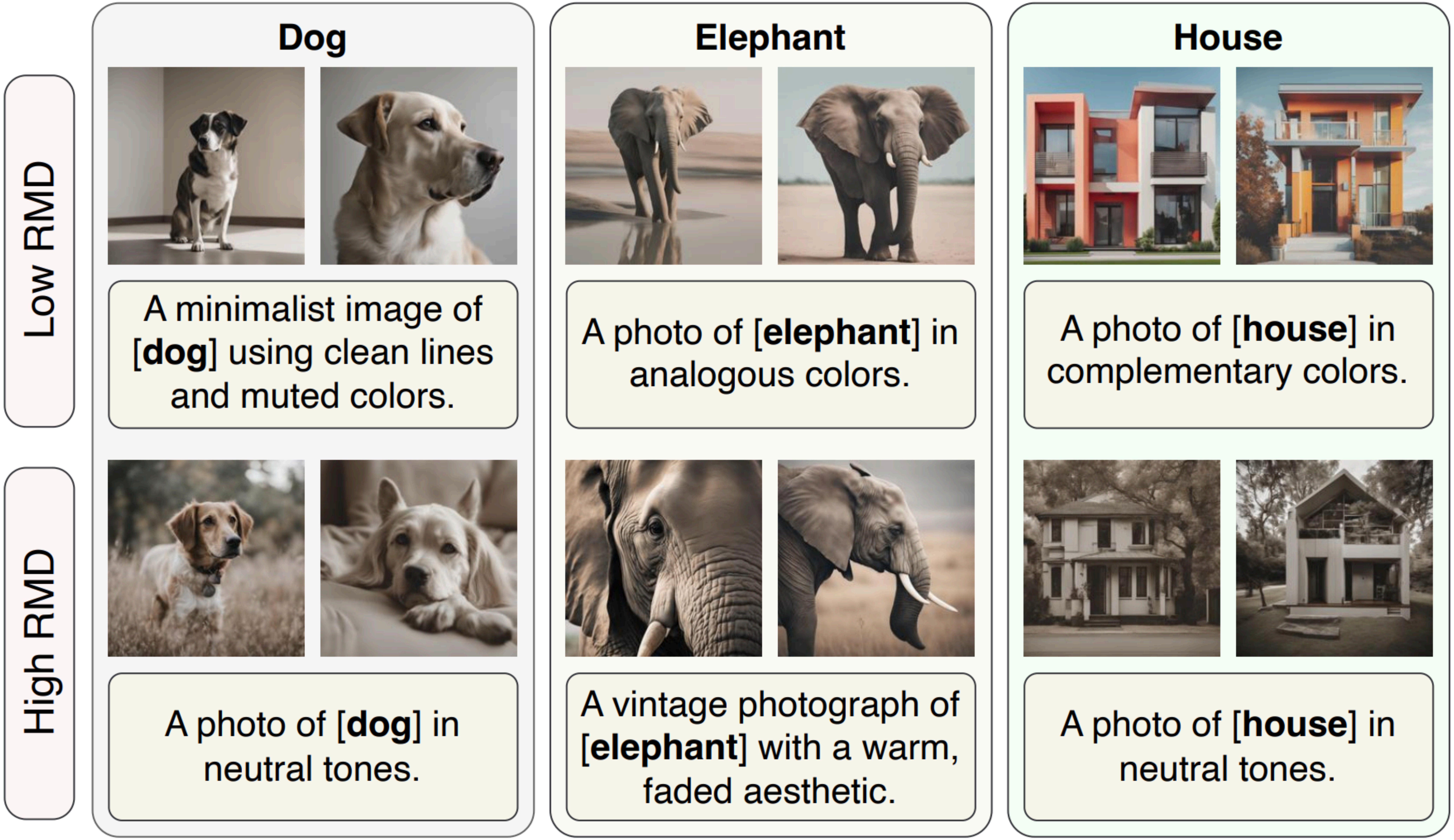


Meta Prompts
A photo of [person] in earth tones.
A vintage photograph of [person] with a warm, faded aesthetic.
Prompt Rewrites
Image of [concept] with a warm and inviting color palette reminiscent of nature, using earth tones.
[concept] photo with a grounded and organic color scheme inspired by the natural world, using earth tones.
Photo of [concept] showcasing a calming and natural color palette with earth tones.
Generate an image of [concept] in the style of a vintage photograph, featuring a warm color palette and faded appearance.
Depict [concept] in a photo reminiscent of old times, with a warm, faded aesthetic and a vintage feel.
Produce a photo of [concept] with a classic aesthetic, using a warm color scheme and a subtle vintage fade.

Sample Complexity as a Measure


Guitar
Giraffe
Elephant
Dog
Low RMD
RMD 0.43096 -4.59762 -1.21591 -16.2159
High RMD
RMD 6.23886 6.32051 6.25181 6.1329
SDXL
DALL.E-2
DeepFloyd-IF
CogView2
RMD score
100
60
20
−20
−60
−100
Eyeglasses
Candle
Tree
Concept Name

Meta prompt	Average RMD score
A black and white image of [concept] highlighting dramatic contrasts.	-3.471
A minimalist image of [concept] using clean lines and muted colors.	-1.153
A photo of [concept] in analogous colors.	-0.618
A photo of [concept] in complementary colors.	-1.216
A photo of [concept] in earth tones.	1.568
A photo of [concept] in neutral tones.	1.779
This is an image of the [concept].	0.492
A realistic image of [concept].	1.203
A vintage photograph of [concept] with a warm, faded aesthetic.	2.425
A high-resolution photo of [concept] capturing fine details.	-0.446


Dog
Elephant
House
Low RMD
High RMD
A minimalist image of [dog] using clean lines and muted colors.
A photo of [elephant] in analogous colors.
A photo of [house] in complementary colors.
A photo of [dog] in neutral tones.
A vintage photograph of [elephant] with a warm, faded aesthetic.
A photo of [house] in neutral tones.

DISCOBER

$$\mathcal{RMD}(x_i, y_i) = \mathcal{M}(x_i, y_i) - \mathcal{M}_{\text{agn}}(x_i),$$

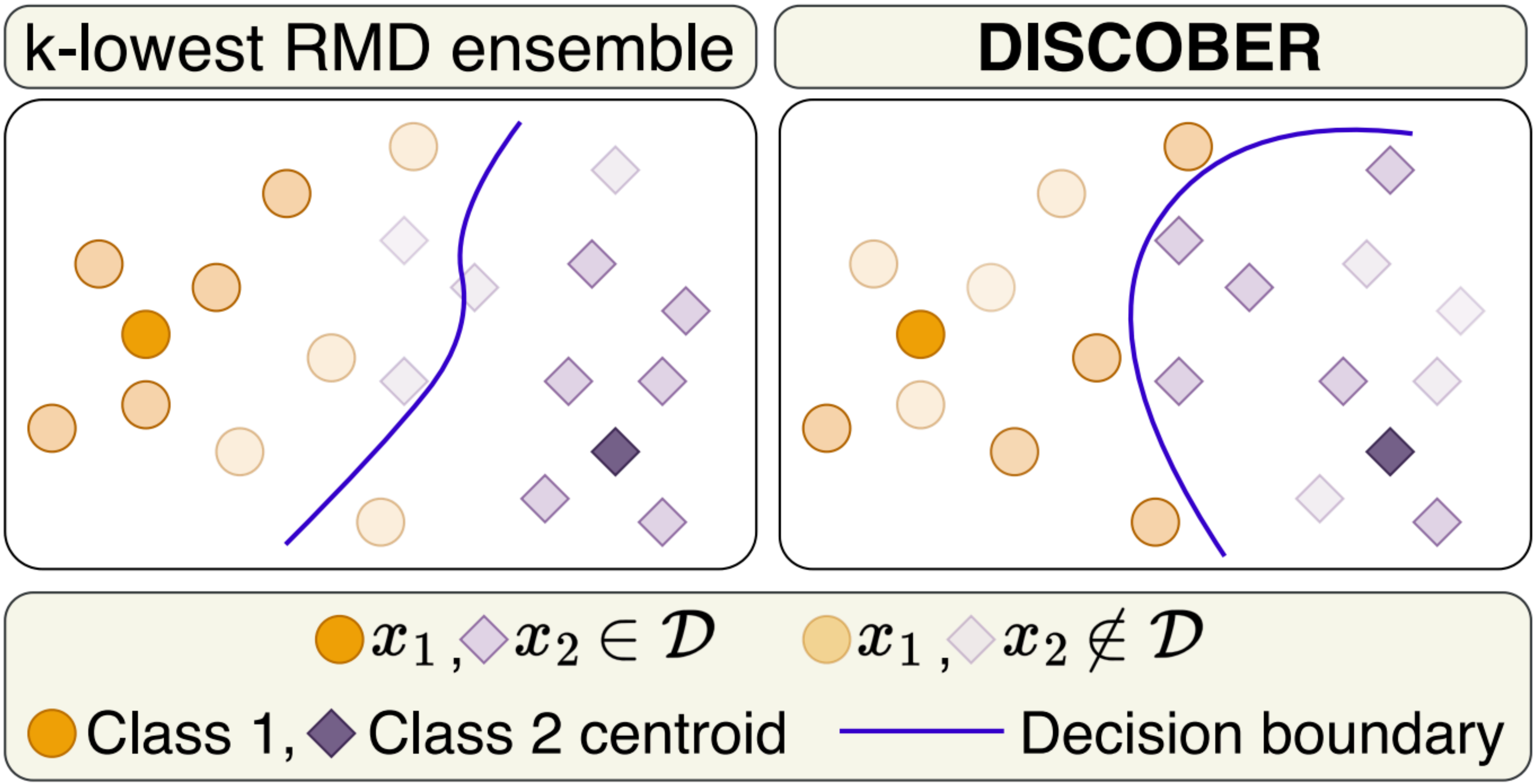
$$p_{g|c} = \frac{e^{\overline{RMD}_{g|c}/T}}{\sum_{h \in \mathcal{G}} e^{\overline{RMD}_{h|c}/T}},$$

High RMD refers to harder samples as measured by distance from global, class prototype

Table 1: Comparison of ensemble methods in PACS [144], using DER [18] for all ensemble methods. The proposed ensemble method outperforms other ensemble methods.

Ensemble Method Δ	ID		OOD	
	$A_{\text{AUC}} \uparrow$	$A_{last} \uparrow$	$A_{\text{AUC}} \uparrow$	$A_{last} \uparrow$
None (Baseline)	47.34±2.64	44.64±3.08	31.33±1.71	25.36±1.31
Equal weight ensemble	43.39±2.01	36.32±2.76	29.77±1.74	21.47±1.73
k-highest RMD ensemble	50.13±1.99	41.60±3.79	31.28±1.23	26.66±1.46
k-lowest RMD ensemble	31.16±0.87	21.60±2.66	25.45±1.56	11.95±1.33
Inverse Prob	40.48±1.72	23.74±0.97	27.98±0.91	20.13±1.37
DISCOBER (Ours)	**50.22±2.41**	**45.10±1.69**	**32.77±1.62**	**28.78±1.49**

DISCOBER interpretation from SVM perspective


k-lowest RMD ensemble
DISCOBER
x_1, $x_2 \in \mathcal{D}$
x_1, $x_2 \notin \mathcal{D}$
Class 1, Class 2 centroid
Decision boundary

Results

Table 2: Split of in-distribution (ID) domain and out-of-distribution (OOD) domain for each domain generalization benchmark.

Dataset	ID domain	OOD domain
PACS [144]	Photo	Art, Cartoon, Sketch
DomainNet [83]	Real	Clipart, Painting, Sketch
CIFAR-10-W [118]	-	CIFAR-10-W [118]
CCT [12]	10 locations	10 other locations

Table 6: Task configurations for class-IL setup on each domain generalization dataset.

Dataset	total # of classes	# of tasks	# of classes / task
PACS [144]	7	3	2 (only initial task: 3)
DomainNet [83]	345	5	69
CIFAR-10-W [118]	10	5	2
CCT [12]	12	4	3

Method	Training Data	PACS				DomainNet			
		ID		OOD		ID		OOD	
		A_{AUC} ↑	A_{last} ↑	A_{AUC} ↑	A_{last} ↑	A_{AUC} ↑	A_{last} ↑	A_{AUC} ↑	A_{last} ↑
ER [99]	Web-scraped	53.08±2.73	50.91±2.57	29.01±2.17	24.70±0.83	**31.98±0.38**	23.29±0.22	9.97±0.23	6.97±0.13
	Base Prompt	46.33±1.75	45.34±3.60	27.96±1.69	20.47±1.39	25.13±0.38	21.38±0.71	7.28±0.15	5.29±0.13
	(+) Diversified Prompt	47.95±2.20	45.58±3.00	34.11±1.33	27.13±1.69	25.23±0.31	20.72±0.35	9.15±0.26	7.35±0.04
	(+) Gen. Ensemble	**53.83±2.96**	**51.68±2.68**	**35.69±1.62**	**30.09±1.42**	28.52±0.07	**24.02±0.86**	**11.42±0.04**	**9.67±0.47**
	Manually Annotated	70.21±3.71	72.11±1.57	28.53±1.81	22.08±1.31	48.56±0.23	40.22±0.55	12.68±0.10	10.19±0.18
ER-MIR [3]	Web-scraped	47.45±4.47	44.57±5.26	27.97±2.20	18.17±1.55	**32.39±0.31**	23.36±0.32	10.25±0.23	7.26±0.07
	Base Prompt	49.34±2.11	46.71±0.83	28.24±1.56	21.00±2.16	24.81±0.43	21.17±0.32	7.23±0.22	5.73±0.15
	(+) Diversified Prompt	50.46±2.18	49.62±3.43	34.36±1.82	28.02±1.16	24.82±0.20	20.56±0.35	9.10±0.20	7.51±0.15
	(+) Gen. Ensemble	**54.28±3.84**	**55.31±1.05**	**37.42±1.80**	**33.90±0.93**	28.36±0.13	**23.74±0.37**	**11.43±0.10**	**9.59±0.19**
	Manually Annotated	68.15±5.06	70.98±1.98	28.78±2.26	21.14±1.04	49.20±0.10	40.54±0.46	12.96±0.03	10.33±0.25
DER++ [18]	Web-scraped	48.39±3.17	36.50±4.24	26.89±1.86	18.88±1.00	**32.09±0.36**	22.37±0.42	9.92±0.20	6.42±0.04
	Base Prompt	41.47±2.26	39.41±2.90	27.74±1.41	18.82±1.57	26.64±0.39	22.04±0.37	7.91±0.24	5.85±0.05
	(+) Diversified Prompt	47.34±2.64	41.60±4.08	32.33±1.71	25.36±1.31	25.61±0.36	20.06±0.38	9.40±0.13	7.20±0.17
	(+) Gen. Ensemble	**49.02±2.41**	**45.10±1.69**	**33.07±1.62**	**28.78±1.49**	29.67±0.06	**23.37±0.38**	**11.89±0.02**	**9.41±0.16**
	Manually Annotated	63.90±5.04	61.19±2.92	27.49±1.77	19.75±1.58	49.35±0.33	39.40±0.20	12.62±0.13	9.27±0.18
ASER [109]	Web-scraped	**49.12±3.32**	42.49±4.06	27.50±1.92	19.04±1.48	**33.80±0.38**	23.09±0.84	9.80±0.51	6.43±0.69
	Base Prompt	40.35±1.25	38.04±2.79	26.64±1.28	18.06±0.80	25.42±0.24	22.93±0.19	7.71±0.64	5.13±0.76
	(+) Diversified Prompt	48.28±0.67	45.40±2.95	33.76±1.20	25.48±1.94	25.94±0.26	20.93±0.31	9.87±0.02	5.64±0.44
	(+) Gen. Ensemble	48.38±1.95	**47.24±2.07**	**35.07±1.46**	**31.58±2.09**	32.01±0.85	**24.28±0.70**	**11.56±0.62**	**8.25±0.98**
	Manually Annotated	68.00±4.95	70.33±2.58	26.81±1.72	19.21±1.16	48.92±0.43	40.93±0.12	10.51±1.27	6.43±0.12
MEMO [143]	Web-scraped	49.27±2.52	39.88±4.93	28.00±1.53	19.19±1.36	**30.17±0.25**	21.40±0.24	9.29±0.27	6.28±0.03
	Base Prompt	43.67±0.90	39.76±4.72	27.22±1.09	17.00±0.67	23.54±0.32	19.45±0.22	6.82±0.16	4.98±0.05
	(+) Diversified Prompt	48.80±1.69	46.59±2.50	32.21±1.55	24.56±0.47	23.59±0.22	19.30±0.30	8.63±0.11	6.83±0.11
	(+) Gen. Ensemble	**50.20±2.37**	**48.72±0.91**	**33.50±1.36**	**29.43±2.79**	26.88±0.35	**21.67±0.20**	**10.61±0.13**	**8.58±0.19**
	Manually Annotated	67.37±4.67	66.94±2.26	27.73±1.59	20.63±0.71	47.04±0.43	38.25±0.45	11.77±0.20	8.99±0.26
X-DER [16]	Web-scraped	50.44±2.93	41.96±2.11	27.57±1.78	20.73±1.06	31.68±0.21	23.00±0.95	10.93±0.44	8.54±0.10
	Base Prompt	44.78±2.77	46.59±2.62	29.86±1.63	22.86±0.99	27.41±0.23	24.11±0.85	7.91±0.65	6.65±0.12
	(+) Diversified Prompt	49.68±2.97	46.94±3.53	33.61±2.07	24.74±2.70	26.72±0.75	21.71±0.43	9.28±0.86	7.65±0.39
	(+) Gen. Ensemble	**50.52±1.57**	**48.19±2.47**	**33.69±1.36**	**26.73±0.54**	**32.14±0.52**	**25.48±0.16**	**12.39±0.74**	**10.04±0.54**
	Manually Annotated	66.19±4.78	68.49±1.85	28.61±1.92	20.54±0.81	50.35±0.20	42.41±0.14	12.99±0.29	10.68±0.83
LiDER [15]	Web-scraped	51.07±3.06	44.69±2.22	27.95±1.60	22.16±1.22	**30.95±0.34**	23.55±0.28	9.93±0.20	7.25±0.08
	Base Prompt	45.73±2.65	43.26±4.86	29.24±1.30	22.12±1.07	24.27±0.20	21.29±0.45	7.05±0.08	5.55±0.06
	(+) Diversified Prompt	51.74±2.48	51.40±2.79	34.04±1.90	27.10±1.41	24.55±0.10	20.78±0.39	9.05±0.16	7.56±0.14
	(+) Gen. Ensemble	**52.46±3.11**	**52.35±3.26**	**36.18±1.44**	**30.94±1.24**	30.09±0.41	**24.04±0.32**	**11.42±0.34**	**9.26±0.29**
	Manually Annotated	66.31±5.69	66.59±2.60	29.11±2.19	21.21±1.03	47.75±0.16	40.06±0.35	12.34±0.09	10.06±0.08

ResNet-18

Method	Training Data	PACS				CCT			
		ID		OOD		ID		OOD	
		A_{AUC} ↑	A_{last} ↑	A_{AUC} ↑	A_{last} ↑	A_{AUC} ↑	A_{last} ↑	A_{AUC} ↑	A_{last} ↑
ER [99]	Web-scraped	47.12±4.67	30.51±5.98	29.78±1.90	15.71±1.94	24.98±1.02	11.00±0.90	21.71±0.75	9.93±0.78
	DISCOBER	**55.25±4.11**	**48.84±3.95**	**33.24±1.62**	**23.14±1.21**	**25.50±0.99**	**12.03±0.81**	**25.16±0.56**	**14.13±0.95**
	Manually Annotated	72.93±5.29	70.51±1.75	30.68±1.95	20.85±0.84	52.20±2.52	34.07±3.41	42.29±1.55	22.10±2.13
ER-MIR [3]	Web-scraped	48.78±5.96	40.95±5.92	28.71±2.24	20.03±3.24	23.07±3.31	12.37±2.78	22.64±2.43	12.20±4.23
	DISCOBER	**50.74±4.09**	**51.51±1.83**	**31.84±1.93**	**25.17±1.05**	**23.72±0.18**	**12.59±0.65**	**24.82±0.34**	**14.01±4.83**
	Manually Annotated	68.21±6.44	73.29±1.90	28.69±1.96	23.03±0.85	37.75±1.36	18.99±1.43	33.38±0.70	15.31±1.27
DER++ [18]	Web-scraped	**53.61±3.39**	**45.71±4.20**	27.66±1.46	18.75±1.63	23.19±0.51	9.17±1.11	22.17±0.60	8.93±0.66
	DISCOBER	50.44±4.32	43.96±3.32	**30.30±1.81**	**20.91±0.86**	**25.24±1.28**	**10.63±0.85**	**24.39±0.92**	**10.17±0.73**
	Manually Annotated	64.81±6.75	61.36±2.37	28.94±2.03	19.95±1.64	44.05±2.67	19.50±2.78	38.02±1.18	17.10±2.21
ASER [109]	Web-scraped	**56.32±5.10**	49.55±4.53	30.67±2.58	21.82±2.04	25.48±1.05	12.84±1.40	22.33±0.85	12.23±0.99
	DISCOBER	56.06±4.60	**52.04±3.85**	**33.99±2.02**	**25.81±0.92**	**26.15±1.74**	**13.97±1.04**	**24.85±1.13**	**12.73±1.36**
	Manually Annotated	77.83±7.77	76.48±9.23	43.37±4.28	35.87±7.47	54.28±1.71	47.67±1.85	45.07±1.56	28.07±0.72

ViT

Table 4: Comparison of Manually Annotated (MA) data and DISCOBER on CIFAR-10-W. We use ResNet-18 as the backbone.

Method	Training Data	A_{AUC} ↑	A_{last} ↑
ER	DISCOBER	**60.93±3.92**	**48.20±0.27**
	MA	48.97±0.56	31.27±2.31
ER-MIR	DISCOBER	**58.19±0.86**	**46.01±0.34**
	MA	44.77±0.86	35.01±2.50
DER++	DISCOBER	**53.88±1.22**	**39.53±1.42**
	MA	45.25±0.07	28.75±1.44
ASER	DISCOBER	**54.34±0.66**	**41.88±1.00**
	MA	50.00±0.59	34.86±1.17
MEMO	DISCOBER	**53.59±0.67**	**41.69±0.67**
	MA	45.40±0.56	30.97±2.13
X-DER	DISCOBER	**57.56±0.75**	**45.97±0.17**
	MA	47.14±0.82	33.41±1.34
LiDER	DISCOBER	**57.13±0.29**	**45.41±2.58**
	MA	46.97±0.42	28.79±4.27

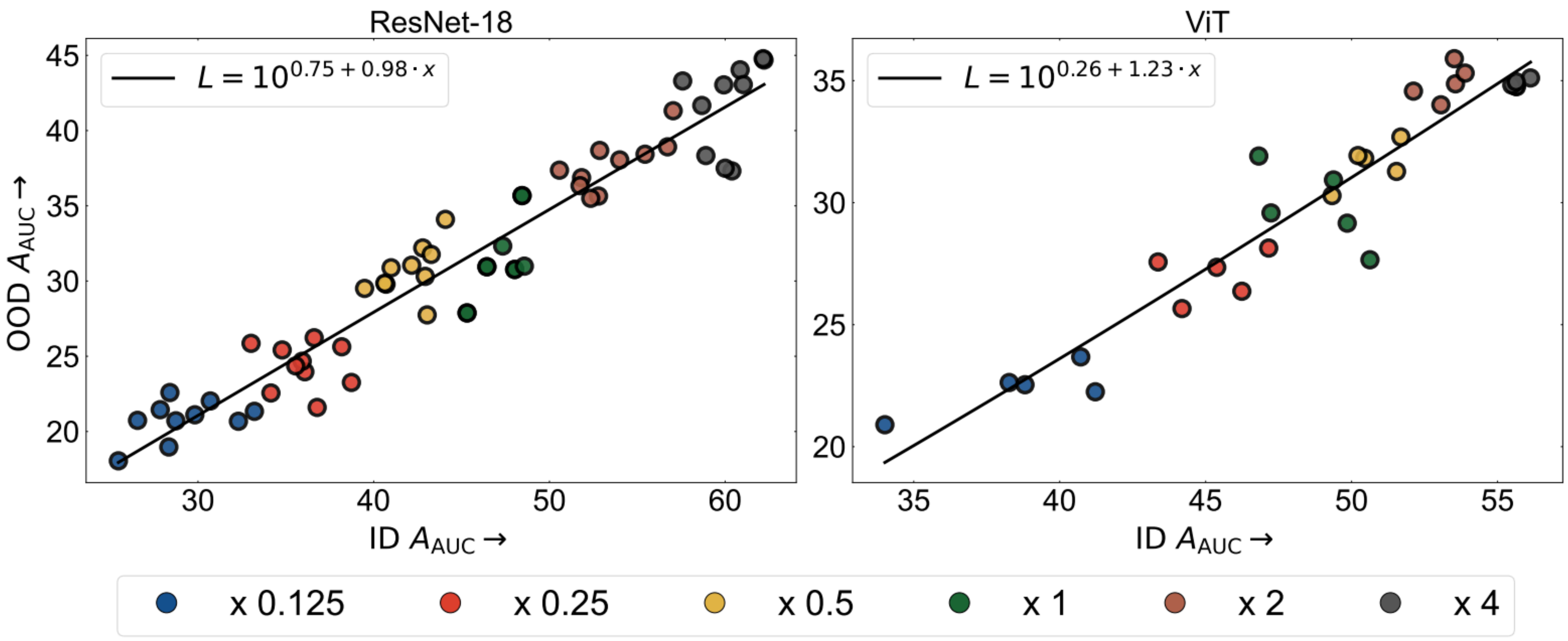


Fig. 5: Ensemble scaling behavior of (a) ResNet-18 [47] and (b) ViT [34] for ID A_{AUC} *vs*. OOD A_{AUC} on the PACS dataset [144] using ER [99]. (`x 1`) denotes the ensemble volume in primary experiments, the default data budget.

Thank You!

digantamisra.github.io @__z__9