Appendices

A. Details on Supervised Training

A.1. Training Hyper-parameters

Supervised training was conducted on both the real ImageNet training set and various synthetic ImageNet generated by text-to-iamge models at different dataset scales. To ensure fair comparisons across different setups, identical training hyper-parameters were used for both real and synthetic images. Our training approach aligns with the setup described in [27], utilizing binary cross-entropy loss. The number of total training iterations and warm-up iterations were adjusted in proportion to the dataset scale in logarithmic space. For instance, models at the 1 million scale were trained for 95k iterations with a 10k iteration warm-up period. At the 2 million scale, training was extended to 190k iterations with a 20k iteration warm-up, and for the 4 million scale, the training and warm-up periods were increased to 285k and 30k iterations, respectively. More detailed descriptions of the training hyper-parameters are provided in Table A1.

Table A1. Detailed pre-training hyper-parameters for supervised training on both real ImageNet training set and synthetic ImageNet generated by text-to-image models.

Config	Value
Batch size	4096
Optimizer	Adam [13]
Learning rate	3×10^{-3}
Weight decay	0.1
Adam β	$\beta_1, \beta_2 = (0.9, 0.999)$
Total iterations	95k for 1M
Warm up iterations	10k for 1M
Learning rate schedule	cosine decay
Mixup	0.5
Dropout	0.1
Stochastic depth	0.1
Augmentation	RandAug(2, 15) [7]

A.2. Details on Text Prompts

In this section, we provide more details on the different configurations of text prompts used for Class-specific Prompts, as outlined in Section 3.1. For the *Classnames + Hypernyms* configuration, we utilized all hypernyms associated with each specific ImageNet category, separated by commas. Regarding *CLIP templates*, we employed two sets of prompt templates with different number of sentences. The first one includes the 80 distinct sentence originally used in the CLIP paper [22] and its inference code[1]. The second set includes a subset of 7 templates, as recommended in [16]. Additionally, we incorporated two more prompt configurations for comparison, following the approach in [25]: (1) *Classnames + Description + Places*, which combines ImageNet class names with their WordNet descriptions, followed by a background category sampled from the Places dataset [32]. (2) *Classnames + Hypernyms + Places*, which is similar to the previous configuration but replaces the descriptions with WordNet hypernyms, also incorporating a background category from Places.

Together with the configurations described in Section 3.1 of the main paper, these methods result in a total of 8 different configurations for text prompts when generating images for ImageNet categories. Additional visualizations of images produced by each these prompt configurations are included in Appendix F.

[1] https://github.com/openai/CLIP/tree/main/notebooks

A.3. Evaluation on Downstream Datasets

In addition to evaluating the trained supervised classifiers directly, we also conducted linear probing on 15 different fine-grained classification datasets. Detailed descriptions of these datasets can be found in Appendix B.2. To perform linear probing on these datasets, we first removed the linear classification head from the classifier trained on ImageNet. Then, we extracted features from both the training and testing sets of each dataset, without applying any data augmentation. Subsequently, logistic regression was employed on these extracted features. The logistic regression layer was optimized using L-BFGS, with a maximum number of iterations equals 500. For a detailed comparison of these results, please refer to Appendix E.

B. Details on CLIP Training

B.1. Hyper-Parameters

Previous studies [19, 22] along with our empirical analysis, indicate the necessity of using different training parameters according to dataset scale. Specifically, for smaller-scale datasets, a larger learning rate and weight decay are recommended to mitigate overfitting. Conversely, for larger datasets, both the learning rate and weight decay should be reduced. Accordingly, we have followed two distinct sets of hyper-parameters within the CLIP training pipeline, one tailored for datasets with fewer than 100 million captions following the parameter in [19], and another for those exceeding this threshold following the parameter in [22]. The specific parameters for both configurations are outlined in Table A3. Models are trained for 32 epochs across all data scales. The number of warmup steps was set to 600 for the 1 million scale, 1200 for the 2 million scale, and 2000 for scales of 4 million or greater. It is important to note that we maintained consistent training hyper-parameters across all three different types of data sources (synthetic, real, syn-

Table A2. Detailed metrics and number of training and testing images of the downstream classification datasets. Only test images are used in the zero-shot classification task for CLIP evaluation.

Dataset	# Categories	# Train Images	# Test Images	Val Metric
Food-101 [1]	101	75,750	25,250	Top-1 Accuracy
CIFAR-10 [15]	10	50,000	10,000	Top-1 Accuracy
CIFAR-100 [15]	100	50,000	10,000	Top-1 Accuracy
SUN397 [30]	397	19,850	19,850	Top-1 Accuracy
Stanford Cars [14]	196	8,144	8,041	Top-1 Accuracy
FGVC Aircraft [18]	100	6,667	3,333	Mean per class
DTD [5]	47	3,760	1,880	Top-1 Accuracy
Oxford Pets [21]	37	3,680	3,669	Mean per class
Caltech-101 [8]	102	3,060	6,085	Mean per class
Oxford Flowers [20]	102	2,040	6,149	Mean per class
STL-10 [6]	10	1,000	8,000	Top-1 Accuracy
EuroSAT [9]	10	10,000	5,000	Top-1 Accuracy
RESISC45 [4]	45	25,200	6,300	Top-1 Accuracy
GTSRB [26]	43	26,640	12,630	Top-1 Accuracy
Country211 [22, 28]	211	42,200	21,100	Top-1 Accuracy

thetic+real) at the same data scale to ensure fair comparisons. For an in-depth comparison of the effects of different hyper-parameters in CLIP training at different scales, please refer to the experimental details provided in Appendix H.1.

Table A3. Detailed pre-training hyper-parameters for CLIP at different dataset scales.

(a) Hyper-parameter for CLIP with $< 100M$ samples.

Config	Value
Batch size	8192
Optimizer	AdamW [17]
Learning rate	1×10^{-3}
Weight decay	0.5
Adam β	$\beta_1, \beta_2 = (0.9, 0.98)$
Adam ϵ	1×10^{-8}
Total epochs	32
Warm up iterations	600, 1200, 2000
Learning rate schedule	cosine decay

(b) Hyper-parameter for CLIP with $\geq 100M$ samples.

Config	Value
Batch size	32768
Optimizer	AdamW [17]
Learning rate	5×10^{-4}
Weight decay	0.2
Adam β	$\beta_1, \beta_2 = (0.9, 0.98)$
Adam ϵ	1×10^{-6}
Total epochs	32
Warm up iterations	2000
Learning rate schedule	cosine decay

B.2. Downstream Dataset

For all the pre-trained CLIP models, we conducted zero-shot evaluations on ImageNet and 15 other widely used downstream classification datasets. These datasets include Food-101 [1], Stanford Cars [14], SUN397 [30], Oxford Pets [21], among others. Detailed information about these evaluation datasets can be found in Table A2. It's important to note that for zero-shot evaluations, only the test images from these datasets are used.

B.3. Zero-shot Evaluation Details

We employed the same text prompt templates as referenced in [22], following a similar text ensembling strategy. For each category, text features were computed for every single template, and the mean average of these features across all templates was used to represent the final text feature for that specific category. Given that CLIP training involves a trainable temperature parameter, τ, it is necessary to incorporate this parameter during zero-shot evaluation to accurately compute the zero-shot classification loss. Let z_{img} be the image feature from the visual encoder, z_{txt} denote the aggregated text feature. Assuming a total of C classes, with z_{txt_c} as the text feature for $c-th$ category, the zero-shot classification loss is calculated as follows:

$$L = -\log \frac{\exp(\text{sim}(z_{img}, z_{txt_c}) \cdot \tau)}{\sum_{c'=1}^{C} \exp(\text{sim}(z_{img}, z_{txt_{c'}}) \cdot \tau)}$$

Here $\text{sim}(z_{img}, z_{txt_c})$ calculates the dot product, measuring the similarity between image feature and text features for each category.

C. Details for Performance at 1.3M

As detailed in Section 4.2 of the main paper, we adopted 54 different configurations encompassing distinct text-to-image models, CFG scales, and text prompts to generate 1.3 million synthetic images for each configuration. Following the generation process, a supervised model was trained on the images geenrated by each configuration. To facilitate a

Table A4. Detailed comparison on ImageNet Validation performance and recognizability, diversity, FID and LPIPS for different CFG scale and prompt configurations with Stable Diffusion as the text-to-image model.

Text-to-Image Model: ***Stable Diffusion****, main comparison on Prompt Config*

CFG Scale	Prompt Config	IN loss(↓)	IN Top1	Recognizability	Diversity	FID(↓)	LPIPS(↓)
2	Word2Sen	3.27	38.42	0.315	**0.850**	3.566	0.717
	CLIP Templates (7)	2.63	49.26	0.522	0.781	3.297	0.714
	CLIP Templates (80)	2.76	49.88	0.510	0.790	3.437	0.719
	Classnames+Hypernym	3.28	45.19	0.569	0.713	2.553	0.698
	Classnames+Description	3.27	45.05	**0.615**	0.696	2.678	0.697
	Classnames+Hypernym+Places	3.14	43.91	0.381	0.836	4.441	0.712
	Classnames+Description+Places	3.06	46.18	0.524	0.758	2.890	0.704
	Classnames	3.05	47.82	0.603	0.718	2.589	**0.702**
	IN-Captions	**2.23**	**55.04**	0.573	0.757	**2.450**	0.714
7.5	CLIP Templates (7)	4.17	38.39	0.702	0.650	5.681	0.731
	CLIP Templates (80)	3.86	40.26	0.687	**0.670**	5.619	0.739
	Classnames	4.96	31.21	**0.780**	0.541	4.113	**0.707**
	IN-Captions	**3.56**	**40.38**	0.725	0.632	**3.641**	0.737

Table A5. Detailed comparison on ImageNet Validation performance and recognizability, diversity, FID and LPIPS for different text-to-image models and CFG scales. All configurations use IN-Captions as prompts.

Text Prompt: ***IN-Captions****, main comparison on Text-to-image models*

Text-to-Image Model	CFG Scale	IN loss(↓)	IN Top1	Recognizability	Diversity	FID(↓)	LPIPS(↓)
Stable Diffusion	1.5	2.14	54.66	0.484	**0.800**	**2.403**	**0.710**
	2	2.23	**55.04**	0.573	0.757	2.450	0.714
	3	2.38	54.10	0.655	0.705	2.790	0.722
	4	2.61	51.13	0.690	0.675	3.100	0.728
	6	2.92	46.99	0.717	0.644	3.483	0.734
	7.5	3.56	40.38	0.725	0.632	3.641	0.737
	8	3.47	40.87	0.726	0.629	3.655	0.738
	10	3.42	33.59	**0.730**	0.621	3.723	0.740
Imagen	1	1.84	58.52	0.466	**0.810**	**3.451**	0.713
	1.5	1.78	**61.51**	0.647	0.719	4.546	0.733
	2	1.93	60.58	0.714	0.671	6.867	**0.701**
Muse	0.1	2.05	54.19	0.473	**0.789**	**4.057**	0.755
	0.3	2.08	**54.45**	0.520	0.760	4.616	0.749
	0.5	2.13	54.03	0.554	0.738	5.189	0.745
	1	2.37	51.55	**0.599**	0.700	6.274	**0.734**

clearer comparison in our tables, we have categorized these 54 configurations into three distinct groups, with each group focusing on specific comparative factors:

- The first group exclusively uses *Stable Diffusion* as the text-to-image model. The primary comparison focus here is on the impact of varying **text prompt** configurations.
- The second group standardizes the text prompt configuration to *IN-Caption*. This group's aim is to assess the effects of using different **text-to-image models** and to understand the behavior of CFG scales within each specific model, and to find the optimal CFG scale for each of them.
- The third group also exclusively uses *Stable Diffusion* as the text-to-image model. Here, the comparison emphasis is on the impact of different **CFG** scales under different text prompt configurations.

By grouping the configurations into these three different groups, we aim to provide a more structured and comprehensible analysis. In each of the three groups, we present the detailed validation loss (the negative log loss here is used to plot Figure 2 in the main paper) and top-1 accuracy on ImageNet validation set for models trained with different configurations, all under the scale of 1.3 million images.

Table A4 presents the analysis for the first group. It

Table A6. Detailed comparison on ImageNet Validation performance and recognizability, diversity, FID(↓) and LPIPS(↓) for different text-to-image models and CFG scales. All configurations use Stable Diffusion as the text-to-image model.

Text-to-Image Model: ***Stable Diffusion****, main comparison on CFG Scale*							
Prompt Config	**CFG Scale**	IN loss(↓)	IN Top1	Recognizability	Diversity	FID(↓)	LPIPS(↓)
ClassNames	1.5	2.84	**48.10**	0.499	**0.778**	**2.525**	**0.702**
	2	3.05	47.82	0.603	0.718	2.589	**0.702**
	3	3.41	44.91	0.697	0.644	2.981	0.704
	4	3.80	41.55	0.734	0.602	3.383	0.705
	6	4.82	33.58	0.770	0.559	3.897	0.707
	7.5	4.96	31.21	0.780	0.541	4.113	0.707
	8	5.04	29.78	0.780	0.537	4.167	0.707
	10	5.47	26.13	**0.787**	0.524	4.289	0.707
ClassNames+Description	1.5	3.05	**45.84**	0.523	**0.753**	**2.468**	**0.697**
	2	3.27	45.05	0.615	0.696	2.678	**0.697**
	3	3.87	41.78	0.699	0.627	3.265	0.699
	4	4.14	38.92	0.738	0.588	3.714	0.701
	6	4.58	33.52	**0.762**	0.546	4.223	0.703
ClassNames+Hypernym	1.5	3.06	**45.22**	0.475	**0.770**	**2.463**	**0.698**
	2	3.28	45.19	0.569	0.713	2.553	**0.698**
	3	3.70	42.52	0.652	0.643	2.961	0.700
	4	4.35	38.02	0.687	0.604	3.336	0.701
	6	4.69	33.38	**0.717**	0.561	3.831	0.703
CLIP Templates (80)	1.25	2.58	47.59	0.344	**0.856**	**3.193**	**0.712**
	1.5	2.61	49.00	0.413	0.831	3.200	0.714
	1.75	2.66	49.77	0.468	0.809	3.284	0.717
	2	2.76	**49.88**	0.510	0.790	3.437	0.719
	3	3.00	48.76	0.600	0.740	4.098	0.726
	4	3.27	46.56	0.642	0.711	4.653	0.731
	6	3.70	42.24	0.677	0.681	5.349	0.736
	7.5	3.86	40.26	**0.687**	0.670	5.619	0.739
CLIP Templates (7)	1.25	2.57	47.86	0.350	**0.851**	**3.029**	**0.709**
	1.5	2.63	49.24	0.424	0.824	3.017	0.711
	1.75	2.69	**50.20**	0.478	0.801	3.116	0.712
	2	2.63	49.26	0.522	0.781	3.297	0.714
	3	3.07	48.69	0.617	0.726	4.025	0.720
	4	3.42	45.93	0.658	0.695	4.661	0.724
	6	4.00	41.15	0.692	0.663	5.393	0.729
	7.5	4.17	38.39	**0.702**	0.650	5.681	0.731
IN-Captions	1.5	2.14	54.66	0.484	**0.800**	**2.403**	**0.710**
	2	2.23	**55.04**	0.573	0.757	2.450	0.714
	3	2.38	54.10	0.655	0.705	2.790	0.722
	4	2.61	51.13	0.690	0.675	3.100	0.728
	6	2.92	46.99	0.717	0.644	3.483	0.734
	7.5	3.56	40.38	0.725	0.632	3.641	0.737
	8	3.47	40.87	0.726	0.629	3.655	0.738
	10	3.42	33.59	**0.730**	0.621	3.723	0.740

shows that using **IN-Caption** as the text prompt yields the best performance across both CFG scales of 2 and 7.5. This superior performance is largely attributed to its ability to guide the text-to-image model to generate diverse images while maintaining high recognizability, thereby justifying our choice of IN-Caption for most of our experiments. In the second group, detailed in Table A5, we observe that different text-to-image models exhibit varying optimal CFG scales for image generation to train supervised models. Specifically, Stable Diffusion, Imagen, and Muse reach their optimal performance at CFG scales of 2, 1.5, and 0.3, respectively. These findings validate our decision to employ these specific CFG scales in our later study of scaling behavior for each model. Table A6 covers the third group's

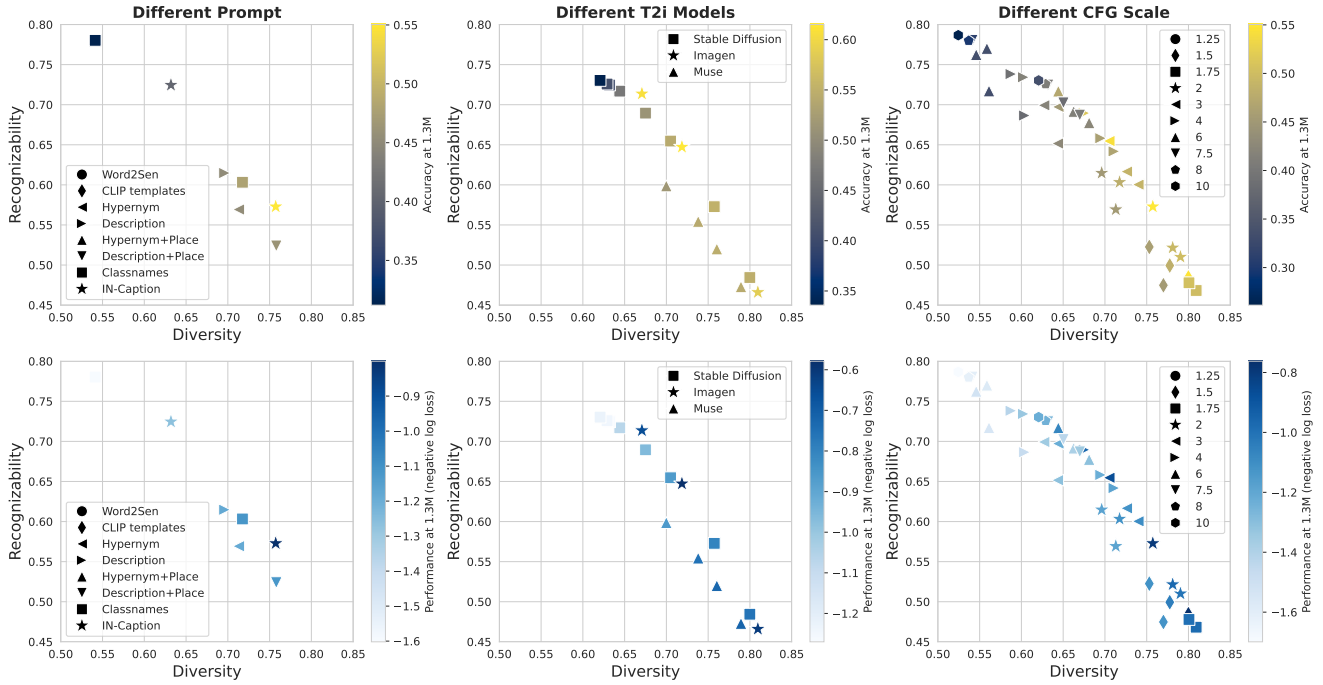


Figure A1. Recognizability versus diversity plot for different text-to-image configuration groups as described in Appendix C. Each column corresponds to one comparison group. The first column mainly compares on text prompts and corresponds to Table A4. The second column compares different text-to-image models and corresponds to Table A5. The last column mainly compares optimal CFG scale for Stable Diffusion and corresponds to Table A6. On the top row, each point is colored by the top-1 accuracy in ImageNet validation set, on the bottom row the points are colored by the negative log loss.

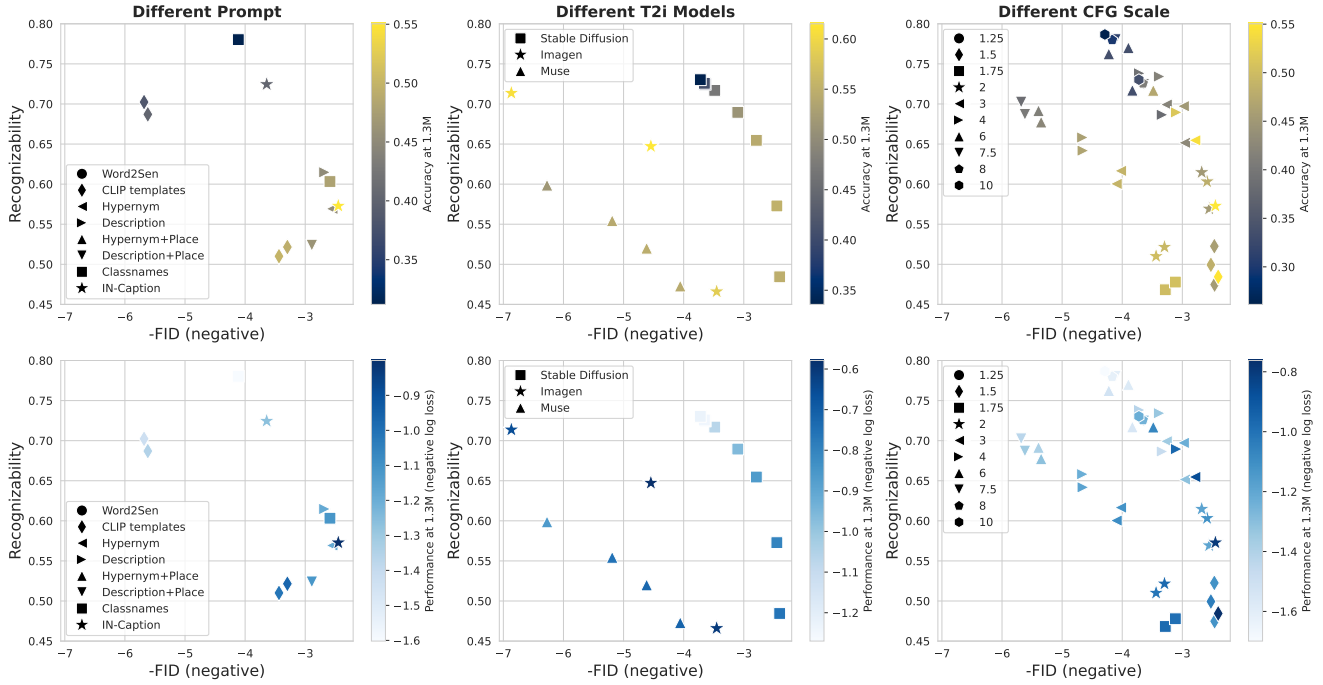


Figure A2. Recognizability versus FID plot for different text-to-image configuration groups as described in Appendix C. Each column corresponds to one comparison group. The detailed correspondence between the plot and tables is the same as Figure A1. On X-axis we take the negative of FID, upper right indicates better metric with lower FID and higher recognizability.

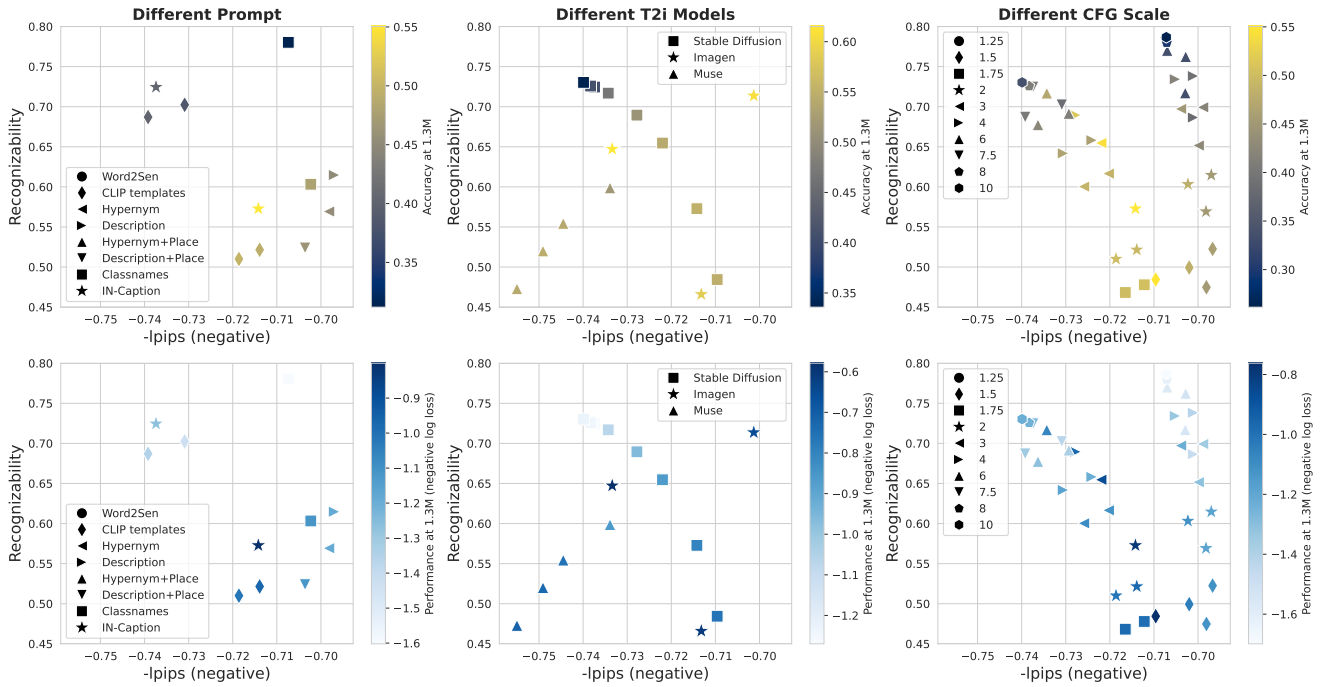


Figure A3. Recognizability versus LPIPS plot for different text-to-image configuration groups as described in Appendix C. Each column corresponds to one comparison group. The detailed correspondence between the plot and tables is the same as Figure A1. On X-axis we take the negative of LPIPS, upper right indicates better metric with lower LPIPS and higher recognizability.

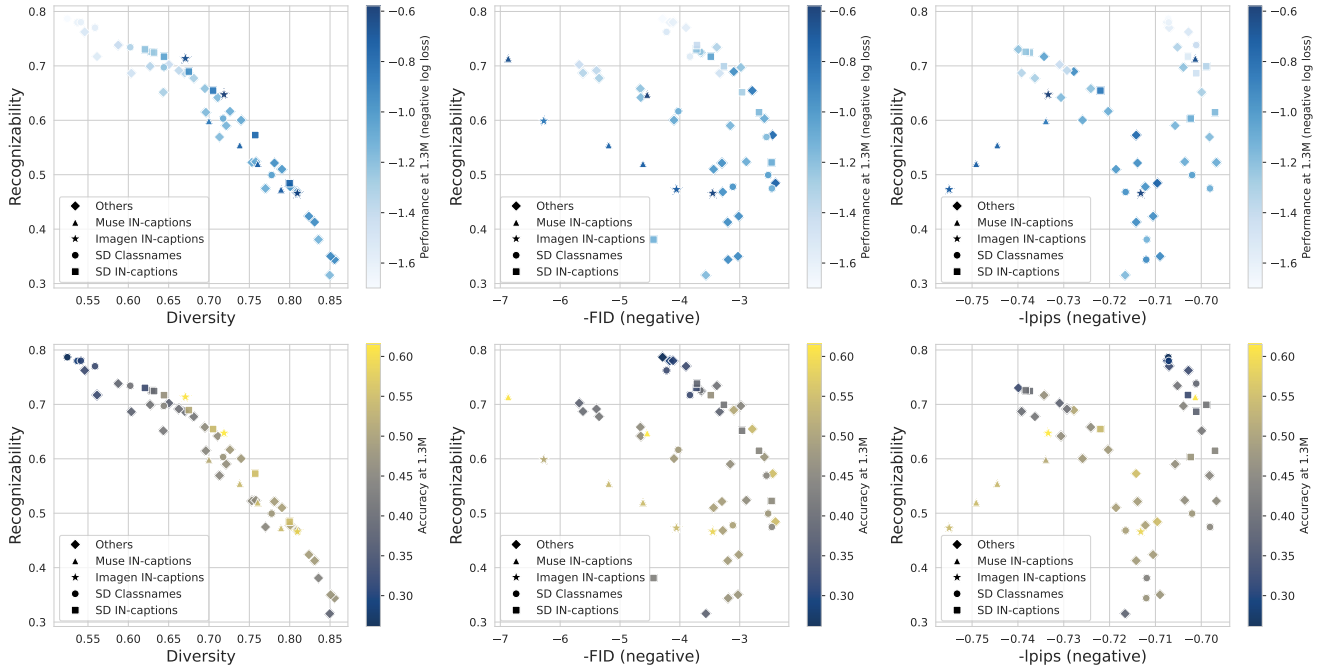


Figure A4. Recognizability versus diversity, FID or LPIPS plot for all the different text-to-image configurations under 1.3M scale. The first column corresponds to diversity, while the latter two correspond to FID and LPIPS, respectively. On X-axis we take the negative of FID and LPIPS. In the plots upper right indicates better metric with higher diversity, lower FID or LPIPS, and higher recognizability. Each point correspond to one configuration, and is color-coded by either the negative log loss (top row) or the top-1 accuracy (bottom) on the ImageNet validation set.

comparisons, focusing on finding the optimal CFG scales for Stable Diffusion under different text prompts. The results indicate that for text prompts with less diversity, such as using Classnames or Classnames+Hypernym, a smaller optimal CFG scale (1.5) is better since it will lead to more diverse images during the generation process. In contrast, for more diverse text prompts, like IN-Captions and CLIP templates with 80 sentences, since there is more diversity on the text side relativly, a larger optimal CFG scale (2) is more effective.

In addition, we have included a recognizability versus diversity plot for each of the three comparison groups in Figure A1. Each point in these plots represents a specific configuration, and is color-coded based on either the top-1 classification accuracy or the negative log loss on ImageNet validation set. The figures illustrate a trade-off between diversity and recognizability. Optimal performance is typically observed when there is a relatively better and more balanced trade-off between these two factors. Configurations characterized by either low diversity or low recognizability tend to result in suboptimal performance, indicating the necessity of maintaining a balance between these two factors.

D. Evaluation under FID and LPIPS

In addition to diversity, we computed two other key metrics: FID (Frechet Inception Distance)[12] and LPIPS (Learned Perceptual Image Patch Similarity)[31]. Both of them are standard evaluation metrics for the text-to-image generation models. Our study examines the performance variations in relation to these two metrics. As detailed in Section 3.2 of the main paper, these metric scores are also calculated using the synthetic test sets, which comprises $50,000$ images for each configuration:

- The FID scores are derived by measuring the Frechet Inception Distance [12] between the synthetic test set, containing these 50,000 generated images, and the real ImageNet validation set.
- For LPIPS, we perform the calculation on a per-class basis. We randomly select and compute the similarity between 250 pairs of synthetic images for each class, and the final LPIPS metric is computed as the average across all classes.

The comparison of FID and LPIPS scores across each group is presented in Tables A4, A5, and A6. Additionally, in Figures A2 and A3, we plot a detailed comparison of the performance across different image generation configuration groups, substituting diversity with either FID or LPIPS. Considering that lower scores for FID indicate better distribution match and for LPIPS implies larger intra-class diversity, we take the negative of these values for plotting purposes. This adjustment ensures consistency with the diversity plot on the X-axis, positioning better results towards the right.

Furthermore, we incorporate comparisons using diversity, FID, or LPIPS as the X-axis for all 54 text-to-image generation configurations in Figure A4. Our findings reveal that while there is a moderate correlation between the FID score or LPIPS of generated images and the classification performance of models trained on them, the relationship is not definitive. In some cases, configurations with the same level of recognizability but lower FID scores or LPIPS show inferior classification performance. This suggests that while FID and LPIPS are effective metrics for evaluating the quality of the images generated by text-to-image models, their correlation with the performance of supervised classifiers trained on synthetic images is not as strong as expected. This observation underscores the need for a more specific metric tailored to evaluate the performance of supervised classifiers trained on such synthetic images.

E. Detailed Scaling Behavior Comparison

In Tables A7 and A8, we present a comparison of the scaling behavior of supervised models trained under various configurations. This comparison is based on linear probing performed on 15 fine-grained classification datasets, as detailed in Appendix A.3. Our findings indicate that, in general, the scaling behavior observed in linear probing on these downstream datasets aligns with the trends seen in the ImageNet validation set. However, there are instances where training on synthetic images surpasses the performance of training on real images, in the Food-101 dataset for example.

Additionally, we have also included the detailed comparison on the out-of-distribution (OOD) validation sets, including ImageNet-A [11], ImageNet-R [10], ImageNet-Sketch [29], and Imagenet-V2 [23]. The results from these comparisons demonstrate that training on synthetic images can yield improved performance on OOD test sets, exemplified by the results on ImageNet-R.

F. Visualization on Generated Images

To better understand the impact of various text prompts used in generating training images, we provide additional visualizations of images created using different text prompt configurations for specific ImageNet categories. These visualizations were generated using Stable Diffusion, with the CFG scale set to 2.

In Figure A6, we present a detailed visualization of the images generated with different text prompt configurations for three different ImageNet categories: Goldfish, Golden Retriever, and shopping carts. The visualizations illustrate that incorporating more detailed information into the prompt tends to encourage the text-to-image model to generate more diverse images. However, this increased diversity may potentially compromise the accuracy of the cate-

Table A7. Detailed scaling behavior on 15 different downstream classification datasets and ImageNet-A, ImageNet-R, ImageNet-Sketch and ImageNet-V2 validation set for supervised classifiers trained with real images from ImageNet training set and synthetic images from various configurations using Stable Diffusion. Dataset scale is in million.

Scale	Food-101	CIFAR-10	CIFAR-100	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	STL-10	EuroSAT	RESISC45	GTSRB	Country211	**DS Average**	ImageNet-A	ImageNet-R	ImageNet-Sketch	ImageNet-V2
Real ImageNet Training set																				
0.125	61.4	80.6	60.9	47.3	24.0	31.0	64.7	77.2	73.4	86.7	88.0	95.8	87.4	57.0	12.0	63.2	2.6	14.6	5.7	36.3
0.25	65.5	85.2	66.2	52.7	30.0	37.7	67.7	83.3	81.0	88.8	92.5	95.8	88.0	61.0	11.9	67.1	3.6	19.6	10.2	45.1
0.5	71.0	89.7	72.5	57.6	44.2	43.4	70.2	88.8	87.2	92.1	96.0	95.8	89.8	65.2	12.4	71.7	5.8	27.3	17.6	55.7
1	77.0	94.7	80.0	63.1	57.3	51.6	72.8	92.6	92.8	93.4	98.1	96.0	90.5	70.8	13.9	76.3	15.6	40.3	29.4	66.7
1.3	77.8	94.6	81.0	64.3	62.5	53.1	74.0	93.5	93.4	93.4	98.6	96.1	90.0	71.7	14.4	77.2	18.7	42.2	31.2	68.8
Stable Diffusion, CFG scale=7.5, Classname																				
0.125	56.4	75.8	54.1	43.2	24.7	31.5	61.9	73.1	65.4	83.9	81.3	94.7	84.6	53.8	9.4	59.6	1.7	15.4	6.8	20.6
0.25	59.1	77.9	56.7	43.9	28.7	35.7	61.1	75.0	67.2	84.5	84.0	94.4	84.7	57.9	9.7	61.4	1.8	16.9	8.1	21.6
0.5	60.2	79.4	58.8	46.5	31.9	37.6	64.5	77.7	73.4	84.7	87.1	95.3	86.3	61.1	10.3	63.7	2.1	19.6	10.2	23.4
1	61.2	82.7	61.4	47.7	35.6	39.4	62.1	79.0	73.0	85.6	87.4	94.9	86.5	62.2	10.3	64.6	2.5	22.5	13.1	25.2
2	61.7	82.6	61.4	48.9	38.5	39.3	63.6	79.4	73.8	85.6	87.9	94.4	86.6	64.1	10.5	65.2	2.6	24.9	15.2	25.8
4	61.5	83.3	62.2	48.6	35.9	40.5	63.2	80.1	76.2	84.7	89.0	94.0	86.1	62.6	10.5	65.2	3.1	26.0	16.3	25.8
8	62.1	83.8	63.4	48.5	38.8	39.7	63.9	79.6	76.4	83.2	89.5	93.8	86.5	63.3	10.6	65.5	2.7	27.3	16.5	26.2
16	61.2	83.1	62.9	49.0	36.3	40.3	62.4	79.4	77.1	83.9	89.4	93.5	86.6	66.5	10.7	65.5	3.3	27.8	18.0	27.2
32	61.2	84.4	63.8	49.6	36.8	38.6	64.0	78.9	76.6	82.8	89.5	93.8	86.0	63.9	10.5	65.4	3.1	28.2	17.9	26.2
64	61.8	83.8	63.7	49.3	37.3	38.5	62.4	80.3	76.9	82.6	89.8	93.8	86.0	64.2	10.7	65.4	3.2	28.9	18.0	26.9
Stable Diffusion, CFG scale=2.0, Classname																				
0.125	61.2	73.2	51.6	46.6	25.6	33.4	62.8	76.0	68.2	85.4	84.1	95.1	86.6	54.9	9.9	61.0	2.4	17.9	6.6	22.3
0.25	65.1	77.5	56.7	51.1	33.6	38.6	66.2	82.3	76.4	88.6	87.4	95.1	88.0	57.2	10.2	64.9	2.9	23.9	10.8	28.7
0.5	69.5	80.7	61.2	54.6	46.9	46.0	68.3	86.9	83.6	90.6	91.5	95.3	89.4	60.6	11.1	69.1	3.3	31.7	16.5	33.8
1	72.5	84.9	65.9	58.8	55.1	50.6	70.8	89.1	87.1	91.9	94.4	96.0	90.1	64.3	12.0	72.2	5.0	41.0	23.0	39.1
2	74.1	86.9	67.5	59.9	55.8	52.1	70.9	88.8	87.8	91.7	95.3	94.7	89.7	69.1	12.3	73.1	6.6	45.7	27.2	42.2
4	75.4	87.1	68.4	60.2	57.2	52.1	71.2	89.0	89.2	91.8	95.8	95.3	89.6	67.3	12.1	73.4	7.9	49.2	29.6	44.3
8	76.3	88.3	69.4	60.8	60.5	53.5	71.8	89.5	89.9	92.2	96.5	94.7	89.8	68.0	12.5	74.2	8.0	50.4	31.0	44.9
16	76.4	88.6	69.9	61.6	59.5	53.9	72.6	88.2	89.2	91.8	96.7	94.5	90.1	67.5	12.8	74.2	8.6	51.8	31.4	45.6
32	76.7	88.8	71.1	61.7	57.0	51.7	72.0	89.7	88.9	92.3	96.6	94.4	89.6	67.7	12.9	74.1	9.0	51.9	32.1	46.1
64	76.6	88.2	69.6	61.7	58.5	53.0	72.0	89.6	89.8	91.9	96.6	95.0	90.1	68.1	12.9	74.2	9.4	52.5	32.4	46.0
Stable Diffusion, CFG scale=2.0, CLIP Templates(80)																				
0.125	60.4	70.4	48.5	45.1	24.9	32.6	63.1	73.0	67.3	86.4	83.1	95.2	86.5	50.9	9.8	59.8	2.4	18.5	7.6	19.6
0.25	64.0	74.5	53.7	49.4	32.2	37.8	65.1	80.5	75.3	88.0	86.7	95.4	87.2	56.4	10.1	63.8	2.7	29.0	14.8	27.4
0.5	67.6	79.6	58.7	54.0	42.7	45.2	68.1	86.8	84.6	90.3	91.0	95.1	88.9	60.5	10.1	68.2	3.1	40.3	24.4	33.5
1	71.7	85.1	65.5	58.6	53.6	50.1	70.6	88.0	88.8	92.6	94.7	96.0	89.3	65.4	11.3	72.1	5.1	52.9	33.5	40.4
2	75.0	87.8	69.3	61.8	61.0	55.1	72.6	90.3	90.9	93.7	96.5	95.4	90.3	68.5	12.0	74.7	7.5	61.9	39.7	45.1
4	76.3	89.7	71.9	62.9	63.3	55.3	74.3	91.8	92.3	94.3	96.6	94.9	91.1	68.9	12.8	75.8	9.5	66.4	42.6	47.4
8	76.9	90.1	71.6	63.0	61.7	56.2	73.9	90.7	91.0	93.5	96.7	95.1	89.6	67.6	12.4	75.3	10.5	67.5	43.8	48.7
16	77.2	89.7	72.1	63.0	63.5	56.0	73.8	90.9	91.3	92.0	97.0	94.5	89.7	67.6	12.4	75.4	10.8	68.8	44.3	49.3
32	77.5	90.2	71.9	62.6	62.5	56.3	73.5	90.8	91.5	93.2	96.6	94.9	89.9	67.8	12.3	75.4	11.4	69.0	44.5	49.2
64	77.5	90.5	73.1	62.8	61.6	55.7	73.5	90.9	91.8	93.0	97.2	94.7	90.1	67.3	12.6	75.5	11.5	69.3	44.7	49.6
Stable Diffusion, CFG scale=2.0, IN Caption																				
0.125	59.5	71.3	49.2	45.7	22.6	31.6	59.5	73.5	64.5	84.3	82.6	95.0	86.1	48.3	9.8	58.9	2.3	13.9	5.0	22.5
0.25	63.1	75.7	54.6	50.3	26.7	36.6	64.6	81.0	73.0	86.9	88.0	95.1	86.7	55.4	10.3	63.2	3.0	19.1	8.1	29.5
0.5	67.6	79.8	58.5	54.1	38.3	41.7	66.6	86.6	80.4	90.0	91.2	95.8	88.4	60.7	10.9	67.4	4.4	25.7	12.8	36.7
1	71.9	84.3	64.2	58.8	48.3	49.1	69.8	89.2	86.4	91.1	94.8	95.5	89.4	60.7	11.9	71.0	7.2	35.7	19.5	45.3
2	76.0	88.0	68.9	62.4	57.5	51.4	72.2	90.5	88.6	93.0	96.6	95.5	90.4	64.5	13.0	73.9	10.1	43.8	25.9	50.7
4	77.2	88.8	69.2	62.5	56.2	54.0	72.0	90.7	89.4	92.8	96.8	95.2	90.0	64.5	13.0	74.2	12.5	48.4	28.6	52.7
8	78.1	88.9	70.4	63.6	56.3	53.7	71.3	90.6	90.5	93.1	97.2	94.7	90.4	66.0	13.4	74.6	14.0	50.5	30.9	54.2
16	78.3	89.5	71.4	63.8	56.4	52.2	72.9	90.8	89.8	92.3	97.3	95.0	89.5	66.0	13.4	74.6	15.7	51.4	30.7	54.8
32	78.7	89.9	71.8	64.1	58.8	53.6	72.0	91.2	90.4	91.8	97.3	95.1	89.8	66.6	13.9	75.0	15.9	52.4	31.5	54.8
64	78.4	89.8	70.9	64.1	58.9	53.4	73.0	91.4	90.7	92.3	97.3	94.5	89.7	67.4	13.7	75.0	16.0	53.2	32.2	55.0

Table A8. Detailed scaling behavior on 15 different downstream classification datasets and ImageNet-A, ImageNet-R, ImageNet-Sketch and ImageNet-V2 validation set for supervised classifiers trained with synthetic images from various configurations using Imagen and Muse. Dataset scale is in million.

Scale	Food-101	CIFAR-10	CIFAR-100	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	STL-10	EuroSAT	RESISC45	GTSRB	Country211	**DS Average**	ImageNet-A	ImageNet-R	ImageNet-Sketch	ImageNet-V2
Imagen, CFG scale=2.0, IN Caption																				
0.125	58.5	71.9	50.7	45.7	22.6	31.5	59.8	78.8	65.7	84.4	83.9	94.9	84.7	52.6	11.0	59.8	2.7	15.6	4.9	29.1
0.25	61.8	78.3	56.8	50.1	28.7	36.3	63.9	84.3	74.6	86.4	89.8	94.6	85.3	54.6	11.1	63.8	3.2	22.2	9.4	35.6
0.5	66.6	81.9	61.1	54.8	38.4	43.4	66.1	88.1	80.1	89.5	93.4	95.4	86.1	58.2	11.3	67.6	4.2	29.0	13.7	42.4
1	71.3	86.3	66.8	59.6	49.5	50.3	69.6	90.9	85.9	90.0	96.0	95.4	88.0	61.3	12.6	71.6	7.9	39.4	19.6	51.1
2	73.8	90.0	70.9	60.2	50.2	52.1	68.0	90.9	87.2	89.7	97.2	95.1	86.3	62.0	12.9	72.4	11.9	45.0	23.2	55.0
4	75.1	90.2	71.8	61.5	49.5	53.7	69.1	91.2	87.8	89.6	97.4	94.7	87.1	62.8	13.0	73.0	14.2	49.7	27.1	57.6
8	75.0	91.2	72.8	62.0	52.4	51.7	67.5	91.7	88.0	88.3	98.1	95.0	86.4	62.1	13.4	73.0	16.4	51.0	27.3	58.7
Muse, CFG scale=2.0, IN Caption																				
0.125	58.3	76.6	54.9	45.3	20.5	29.6	62.0	73.2	66.6	84.8	87.5	95.2	83.3	52.2	10.7	60.0	2.7	16.8	6.8	23.4
0.25	63.3	84.1	63.5	50.2	27.0	35.9	65.3	80.3	76.5	87.5	92.5	95.5	85.3	59.4	11.2	65.2	3.6	25.4	13.2	30.5
0.5	67.0	87.0	68.1	54.0	37.3	42.4	67.7	85.1	82.2	89.1	94.5	95.5	86.4	64.8	11.4	68.8	4.6	32.7	19.8	36.4
1	71.2	91.3	72.8	58.9	48.1	47.6	70.4	87.2	87.0	92.1	96.8	95.4	87.2	66.9	11.8	72.3	7.7	43.3	27.9	44.1
2	73.9	92.0	74.3	60.3	49.7	49.3	70.6	89.4	87.6	91.0	97.1	94.6	87.3	67.9	12.7	73.2	12.2	48.8	32.7	48.4
4	74.9	92.9	74.4	60.7	51.4	51.0	71.1	89.6	88.4	91.5	97.7	94.7	86.9	68.3	12.6	73.7	14.4	52.0	34.8	50.3
8	75.3	91.7	73.3	61.1	51.4	51.9	70.5	89.6	87.7	91.6	98.1	93.5	86.9	66.1	12.8	73.4	15.7	53.2	35.6	51.0

gory of interest in the generated images.

G. More per-class analysis

G.1. Recognizablity Distribution

To delve deeper into how recognizability and diversity are distributed across the 1000 ImageNet classes and their influence on scaling ability (k), we categorized all classes into 10 different groups based on their scaling ability, ranging from lowest to highest. For each group, we calculated the average recognizability and diversity of the classes within it. The result of this analysis is illustrated in a detailed bar plot in Figure A5.

The analysis reveals the following trend: as the scaling ability of a group increases, the average diversity initially rises and then begins to decrease. This trend suggests that at the initial stages, enhanced diversity contributes to the generation of more varied synthetic images, helping the supervised classifier to learn more robust features during training. However, beyond a certain point, further increases in diversity can be harmful, potentially compromising the accuracy of the generated images or leading to the omission of key objects or the generation of wrong concepts.

In contrast, the average recognizability consistently increases as scaling ability increases, indicating a stronger correlation between the scaling ability and recognizability for each class. This consistent improvement shows the significance of recognizability as a more relevant metric for class-based analysis.

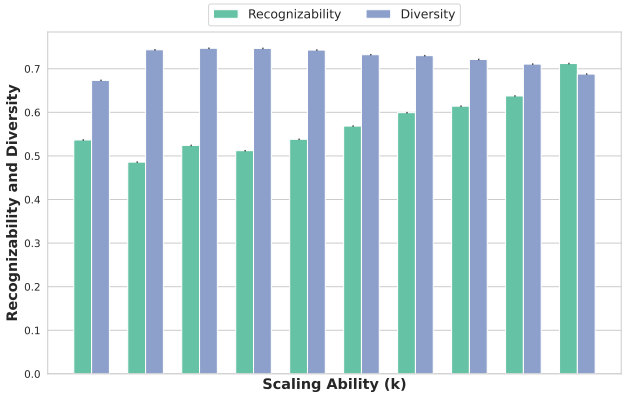


Figure A5. Per class analysis on the changes in recognizability and diversity as the scaling ability (k) increases. Here we divide the 1000 ImageNet classes into 10 groups based on their scaling ability ranging from the lowest to the highest.

G.2. More Results on 'Scaling' Classes

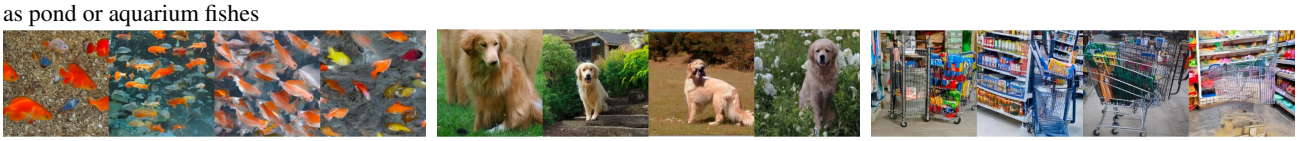
In Figure A7, we provide a detailed comparison of the scaling behavior for models trained on either real or synthetic images from Stable Diffusion, specifically focusing on the 'scaling' classes as described in Section 4.6. Additionally, Figure A8 presents visualizations of synthetic images generated for these classes, using the same setup as described

(a) Classnames

goldfish | Golden Retriever | shopping cart

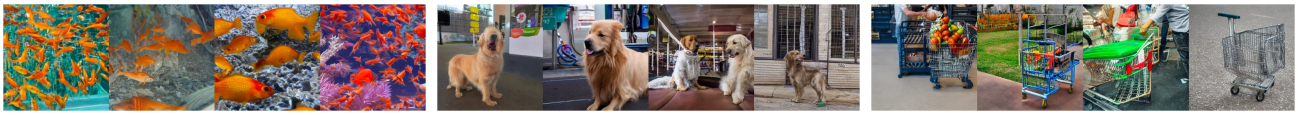
(b) Classname + Description

goldfish, Carassius auratus, small golden or orange-red freshwater fishes of Eurasia used as pond or aquarium fishes | Golden Retriever, an English breed having a long silky golden coat | shopping cart, a handcart that holds groceries or other goods while shopping



(c) Classname + Description + Places

goldfish, Carassius auratus, small golden or orange-red freshwater fishes of Eurasia used as pond or aquarium fishes inside ice skating rink | Golden Retriever, an English breed having a long silky golden coat inside gas station | shopping cart, a handcart that holds groceries or other goods while shopping inside basketball court

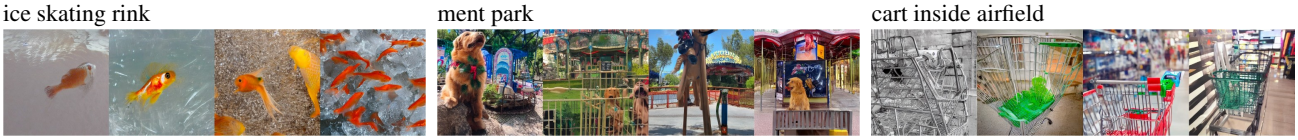


(d) Classname + Hypernym

goldfish, Carassius auratus, cyprinid | Golden Retriever, retriever | shopping cart, handcart, pushcart, cart, go-cart

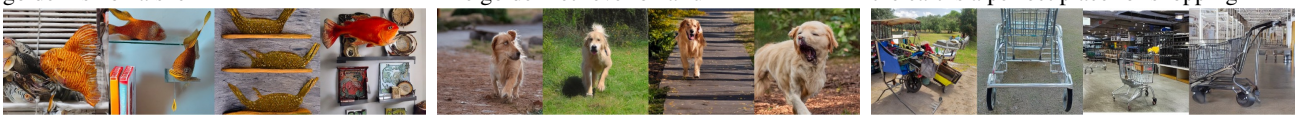
(e) Classname + Hypernym + Places

goldfish, Carassius auratus, cyprinid inside ice skating rink | Golden Retriever, retriever inside amusement park | shopping cart, handcart, pushcart, cart, go-cart inside airfield



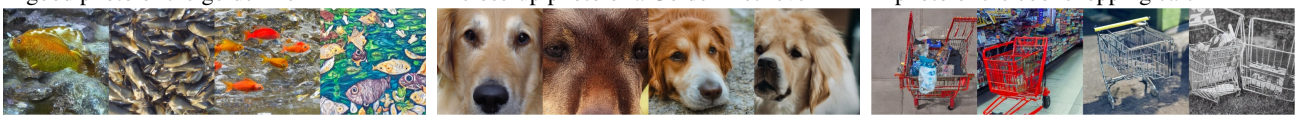
(f) Word2Sen

goldenfish on a shelf | The golden retriever on a run | the cart is a perfect place for shopping



(g) CLIP templates

A good photo of the golden fish | A close-up photo of a Golden Retriever | A photo of the cool shopping cart



(h) IN-Caption

goldfish, fish is swimming in the water | Golden Retriever, a dog is running on the grass | shopping cart, a street with a garbage can and a cart

Figure A6. Synthetic images generated by Stable Diffusion with different text prompt configurations on three ImageNet categories: goldfish, Golden Retriever, and shopping cart. All of these visualizations use a guidance scale of 2.0.

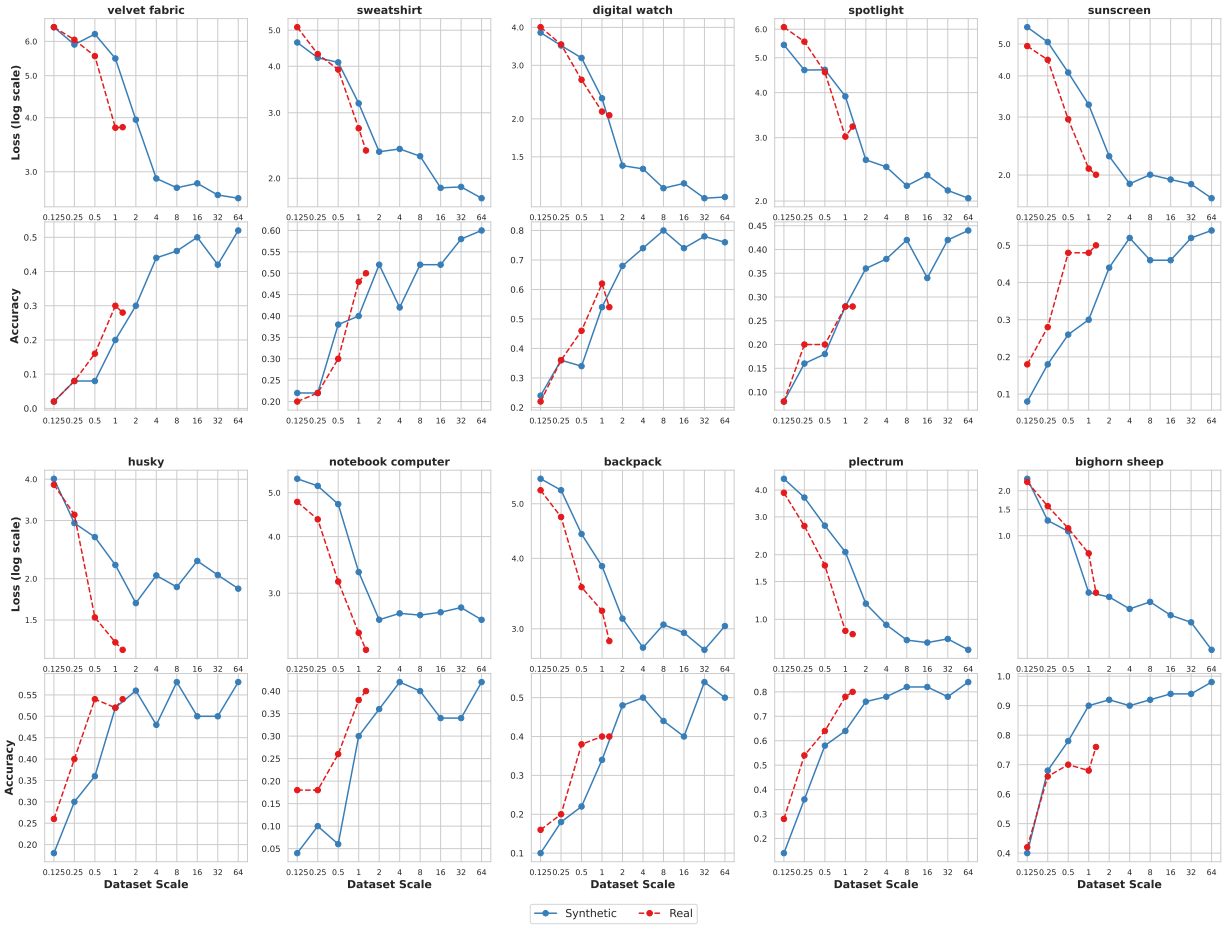


Figure A7. More comparison on supervised models trained on real and synthetic images (from Stable Diffusion), for the 'Scaling' classes.

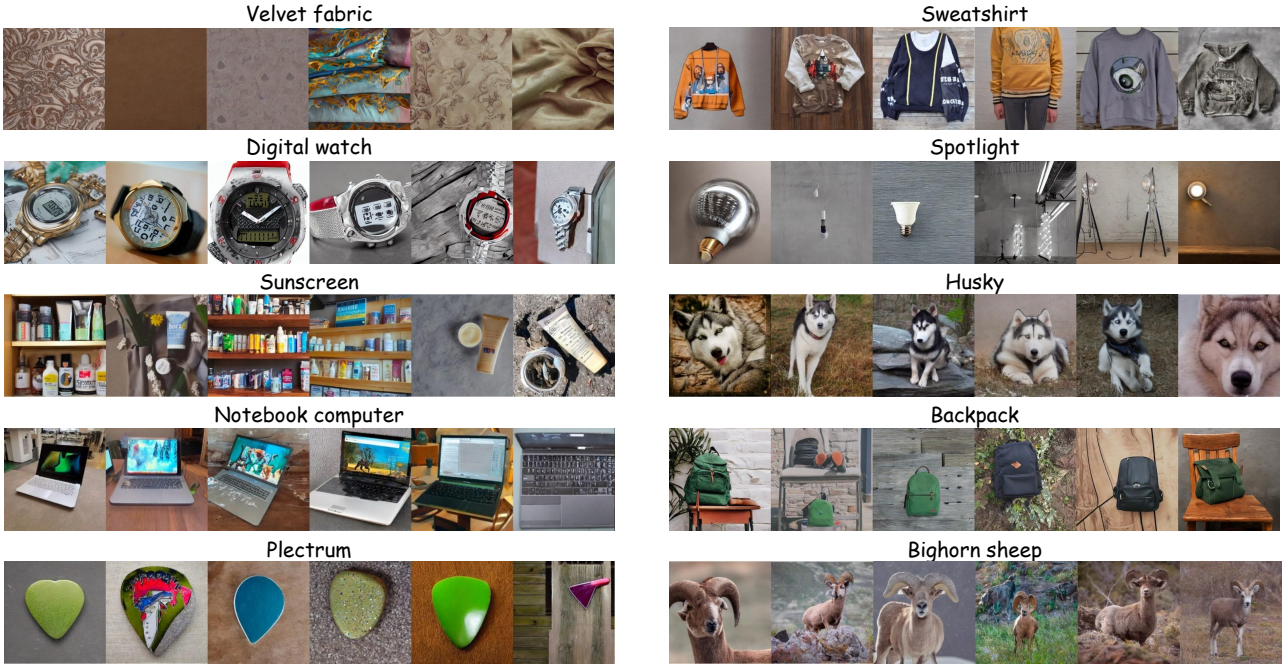


Figure A8. Visualizations of the synthetic images generated for 'Scaling' classes, using Stable Diffusion with a guidance scale of 2.0.

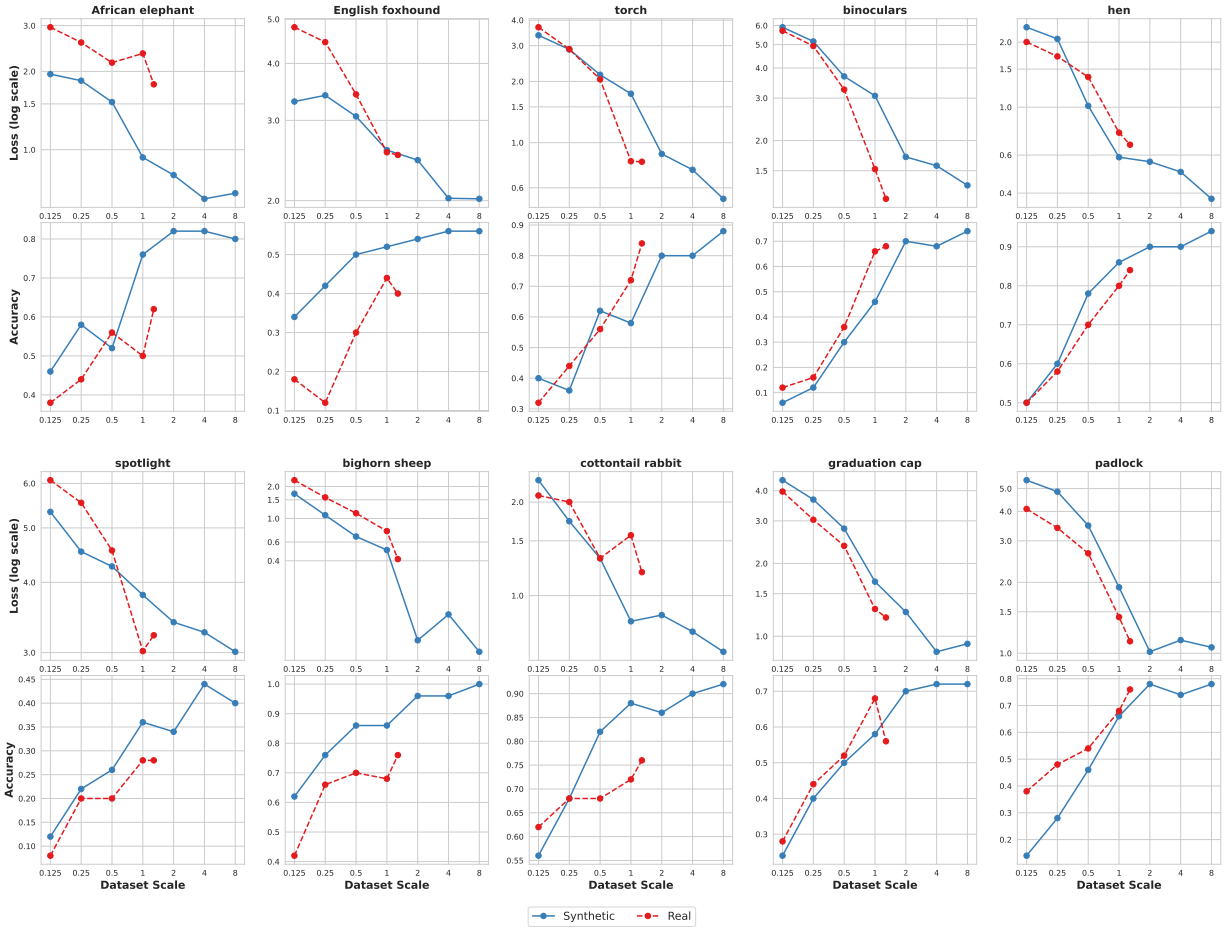


Figure A9. More comparison on supervised models trained on real and synthetic images (from Imagen), for the 'Scaling' classes.

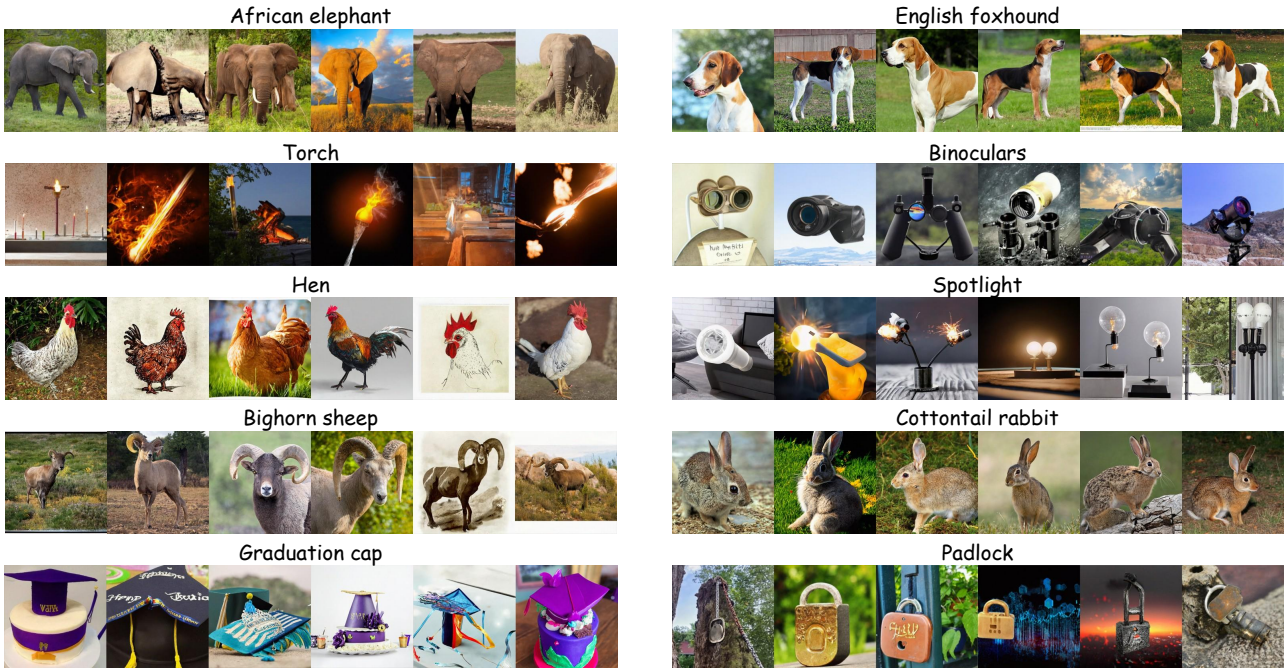


Figure A10. Visualizations of the synthetic images generated for 'Scaling' classes, using Imagen [24] with a guidance scale of 1.5.

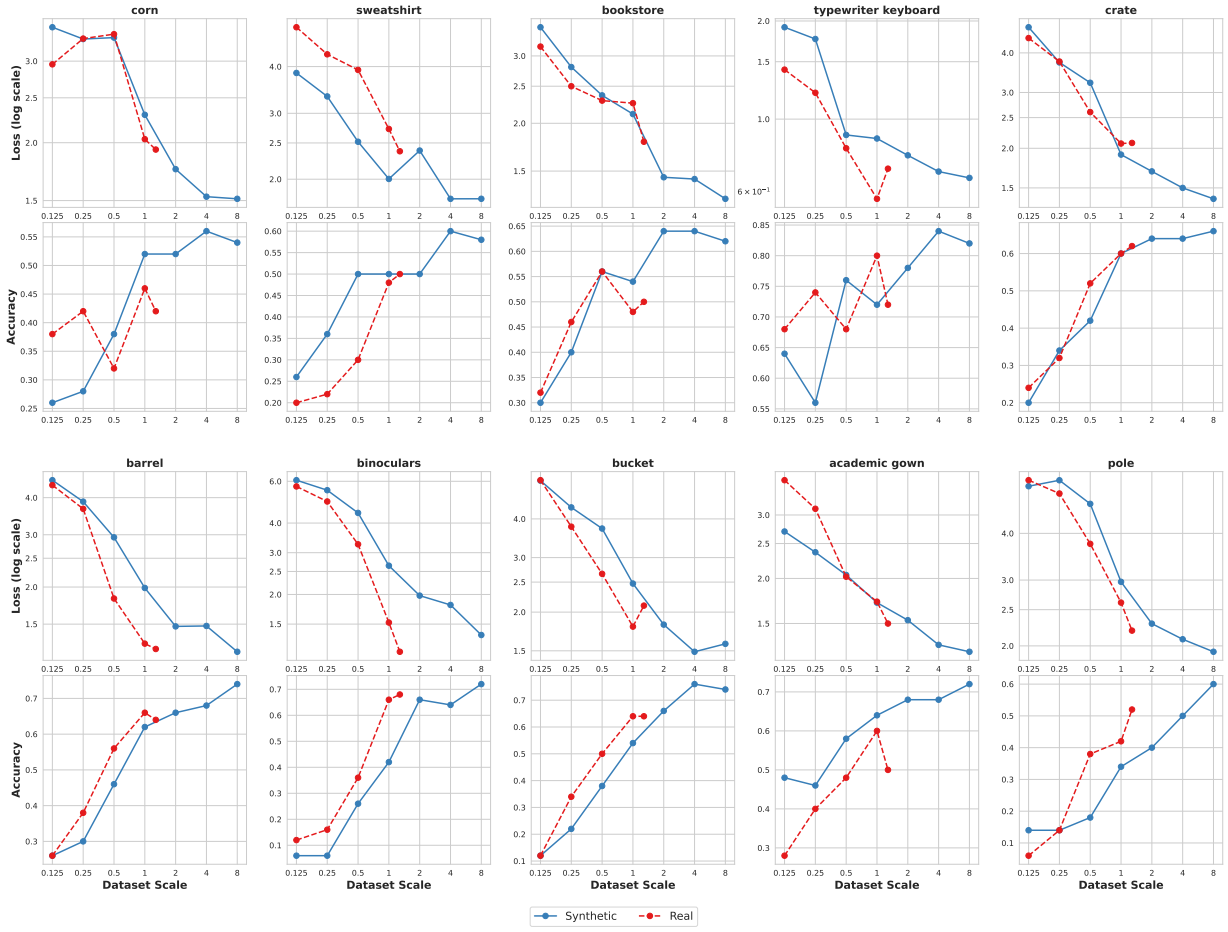


Figure A11. More comparison on supervised models trained on real and synthetic images (from Muse), for the 'Scaling' classes.

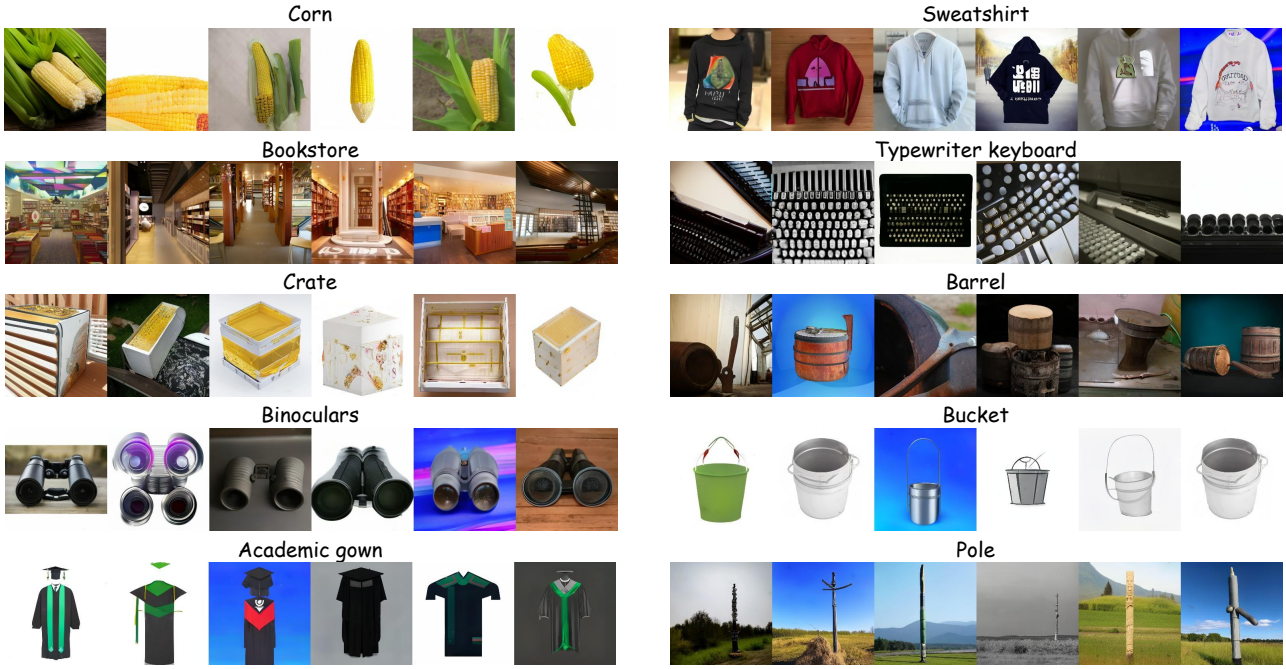


Figure A12. Visualizations of the synthetic images generated for 'Scaling' classes, using Muse [2] with a guidance scale of 0.3.

in Section 4.6.

We further explore 'Scaling' classes for supervised classifiers trained on images generated by the Imagen and Muse models. The scaling behaviors of these classes, in comparison to models trained with real images, along with their visualizations, are presented in Figure A9, Figure A10, Figure A11 and Figure A12. Our analysis reveals that certain classes, such as sweatshirts, exhibit consistently good scaling across different text-to-image models. Meanwhile, there are classes that show particularly strong scaling performance with specific text-to-image models.

For the ten 'scaling' classes selected in Stable Diffusion, we observed that models trained on synthetic images exhibit scaling abilities that are comparable to, and in some cases even superior to, those trained on real images. A notable example can be seen in the 'bighorn sheep' and 'spotlight' categories, where models trained on synthetic images already outperform those trained on real images at dataset scales below 1 million, and this advantage continues to grow as the scale increases, since there are only 1.3M real images.

This finding suggests that for certain concepts, text-to-image models are indeed capable of generating images that are more conducive to train supervised classifiers effectively. As text-to-image models continue to improve, we anticipate that such instances will become more frequent. Eventually, it's plausible that models trained on synthetic images could surpass the performance of those trained on real images across the entire ImageNet validation set.

G.3. More 'Poor' Classes

In Table A9, we identify and list 'poor' classes where supervised models, trained on synthetic images, face challenges in accurate classification. For each of the three text-to-image models — Stable Diffusion, Imagen, and Muse — we highlight 40 distinct categories that pose difficulties. Notably, certain categories, such as tiger cats and vine snakes, are common challenges across different text-to-image models. Future research in the development of text-to-image models could benefit from focusing on these categories. Improving the accuracy in generating images of these 'poor' classes is crucial, as their current limitations are a key factor hindering the ability of synthetic images to have better scaling ability and performance than real images, in the supervised learning contexts.

G.4. What affects Scaling Ability

When we fix the generation configuration of specific text-to-image model, CFG scales and prompts as described in Section 3.1, text-to-image models could still exhibit varying degrees of recognizability and diversity when generating images for different object classes. To explore how these factors influence the scaling ability of each class, we conducted an analysis focusing on the correlation between

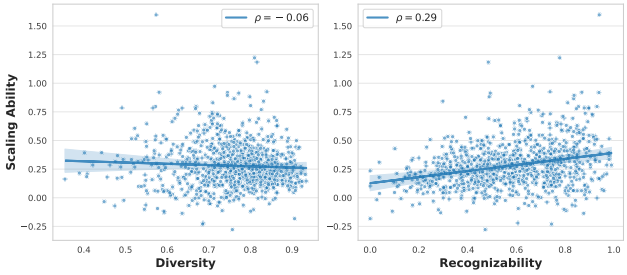


Figure A13. Per class analysis on the relationship between scaling ability (k in Equation 2) and both diversity and recognizability. Within each class, the plots indicate a positive correlation between recognizability and scaling ability. The correlation between diversity and scaling ability appears to be negligible.

scaling ability and both recognizability and diversity, all computed for each class individually. These correlations, and their implications for scaling efficiency, are depicted in Figure A13. Our analysis underscores the potential positive role of recognizability in determining the scaling ability for the synthetic images, for each specific class, generated by text-to-image models. We identified a positive correlation between recognizability and scaling ability, indicating that the precision in generating the intended class significantly enhances the scaling effectiveness of synthetic images. In contrast, the influence of diversity within each class seems to be more limited. Our findings reveal only a negligible correlation between diversity and scaling ability. This might be attributed to the increased noise introduced when computing diversity for specific categories, as opposed to the overall dataset.

G.5. Per-class FID and LPIPS

Following the same setup outlined in Section G.4 of the main paper, we also computed the correlations between scaling ability (k in Equation 2 of the main paper) and both FID and LPIPS scores. Unlike the previous analysis focusing on recognizability and diversity, this evaluation specifically studies the relationship of scaling ability with these two metrics.

To calculate the per-class LPIPS scores, we used the same method as detailed previously. However, for per-class FID computation, the existing synthetic test sets, containing only 50 images per class, were deemed insufficient, since FID score is sensitive to the number of images. Therefore we sample 1300 images from the synthetic training images and compute the FID with images from real ImageNet training set for each class. Similar to our previous approach, we took the negative of the per-class FID and LPIPS scores for consistency, as lower scores indicate better performance. The correlations obtained are depicted in Figure A14.

The results from this figure indicate a lack of strong correlation between scaling ability and either FID or LPIPS

Table A9. Lists of 'poor' classes that has poor scaling ability and performance. Supervised models trained with synthetic images struggles in classifying them correctly. We list 40 categories for Stable Diffusion, Imagen and Muse, respectively.

(a) Stable Diffusion

fire salamander	Appenzeller Sennenhund	tiger cat	collie	Australian Terrier
African bush elephant	cassette player	canoe	European green lizard	night snake
mushroom	eastern hog-nosed snake	hot tub	wall clock	crayfish
espresso machine	water jug	toy terrier	Brittany dog	keyboard space bar
shower curtain	gymnastic horizontal bar	African rock python	letter opener	ladle
tape player	tea cup	paper towel	wok	flute
vine snake	black-footed ferret	cricket insect	European polecat	cradle
Lakeland Terrier	green mamba	cleaver	breastplate	monitor

(b) Imagen

kit fox	shower curtain	night snake	hot tub	minivan
desktop computer	keyboard space bar	European green lizard	espresso machine	black-footed ferret
water jug	flute	velvet fabric	mobile phone	digital clock
product packet / packaging	CRT monitor	eastern hog-nosed snake	tape player	bolete
tobacco shop	monastery	purse	mushroom	printer
letter opener	wall clock	toilet paper	monitor	sunglasses
overskirt	hard disk drive	ladle	can opener	tiger cat
combination lock	paper towel	plunger	tights	vine snake

(c) Muse

titi monkey	alligator lizard	European green lizard	cottontail rabbit	African rock python
stopwatch	gar fish	Irish Water Spaniel	European polecat	CRT monitor
toy terrier	keyboard space bar	night snake	Norfolk Terrier	Ibizan Hound
mobile phone	ground beetle	Tibetan Terrier	Norwich Terrier	purse
Treeing Walker Coonhound	Siberian Husky	eastern hog-nosed snake	Bouvier des Flandres dog	patas monkey
Australian Terrier	CD player	Briard	Affenpinscher	English Setter
cradle	red wolf or maned wolf	Geoffroy's spider monkey	Border Terrier	Lakeland Terrier
tape player	Cairn Terrier	Bluetick Coonhound	Entlebucher Sennenhund	Redbone Coonhound

scores. This finding highlights the necessity for a more tailored metric that is specifically designed to assess the scaling ability of supervised classifiers trained on synthetic images.

H. More Results on CLIP Scaling

H.1. Comparison on Hyper-parameters

In Table A3, we detail the use of two distinct sets of hyper-parameters for CLIP training, tailored to different dataset scales. Config (a) in the table, labeled as 'S' here, is designed for smaller dataset scales with fewer than 100 million images. Conversely, config (b), represented as 'L' here, is intended for larger dataset scales equal to or exceeding 100 million images. To validate the necessity of these configurations, we present an empirical study in Table A10.

Here we train CLIP models on subsets of the LAION-400M dataset with 10M, 50M, and 100M samples, exclusively utilizing real images and applying the two different hyper-parameter sets. Our findings indicate that for scales

Table A10. Comparison of CLIP models trained on LAION-400M subsets of different data scales, using different hyper-parameter configurations. Hyper-parameter configuration 'S' and 'L' corresponds to (a) and (b) in Table A3, respectively. All models are trained on real images and text only, using ViT-B/32 as backbone architecture.

Scale	Hyper-param	Food-101	CIFAR-10	CIFAR-100	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	STL-10	EuroSAT	RESISC45	GTSRB	Country211	**Average**	**ImageNet**
10M	**S**	42.0	78.3	49.0	36.1	40.3	3.6	22.2	47.4	75.0	29.9	85.0	27.5	30.4	19.6	4.4	**39.4**	**30.6**
	L	15.8	51.0	22.7	16.0	9.9	0.8	10.2	19.7	50.2	14.9	63.2	12.5	16.5	6.0	1.9	20.8	14.8
50M	**S**	63.9	87.7	65.4	54.5	61.6	5.9	34.4	71.2	84.0	45.3	93.4	46.4	45.3	28.1	8.4	**53.0**	**47.9**
	L	62.3	84.2	59.9	48.9	62.1	4.9	31.5	71.3	83.2	47.1	90.6	30.9	39.5	30.0	7.7	50.3	44.2
100M	**S**	69.5	88.9	68.6	58.4	67.3	6.9	43.4	75.4	85.7	49.9	93.3	46.9	54.1	41.1	9.6	57.3	51.8
	L	72.6	89.4	68.0	57.6	72.6	7.1	41.0	80.9	87.3	55.9	93.8	36.4	52.3	41.5	10.4	**57.8**	**54.2**

Table A11. Comparison of CLIP models trained on synthetic CC12M generated by Stable Diffusion with different CFG scales. Models are trained using ViT-B/16 as backbone architecture.

CFG	Food-101	CIFAR-10	CIFAR-100	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	STL-10	EuroSAT	RESISC45	GTSRB	Country211	**Average**	**ImageNet**
1.25	41.9	37.8	16.3	39.5	22.8	3.2	20.9	56.3	70.9	22.4	83.8	11.7	31.5	7.2	4.7	31.4	34.9
1.5	43.1	32.8	16.6	42.4	26.6	3.7	23.4	58.8	68.8	24.3	86.4	10.8	33.3	7.4	5.6	32.3	**35.7**
1.75	43.5	29.5	13.3	42.7	28.0	3.9	22.2	58.1	70.6	21.8	84.2	17.7	32.0	8.3	5.5	32.1	35.6
2.5	41.3	47.0	15.1	41.9	28.8	4.7	23.9	57.8	71.2	24.4	87.4	16.8	31.0	6.1	5.7	33.5	34.3

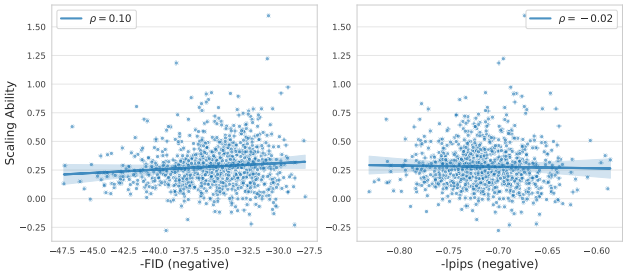


Figure A14. Per class analysis on the relationship between scaling ability (defined as$k$ in Equation 2 in the main paper) and both FID and LPIPS. Within each specific class, the plots indicate the correlation between the scaling ability and both metrics appears to be negligible.

of 10M and 50M, the 'S' hyper-parameter configuration yields superior results, with the performance difference being reduced at the 50M scale. In contrast, at the 100M scale, the 'L' configuration demonstrates enhanced performance. Therefore, based on these empirical results, we opted to utilize the 'S' hyper-parameter set for smaller data scales and the 'L' set for larger scales.

H.2. Comparison on different CFGs

To identify the most effective CFG scale for generating synthetic images to train CLIP models, we utilized the Stable Diffusion to create synthetic images for the CC12M dataset [3] at four different CFG scales: 1.25, 1.5, 1.75, and 2.5. Following the generation of these images, CLIP models were trained using the synthetic images and their corresponding texts. The efficacy of these trained models was then evaluated through zero-shot classification on ImageNet and various downstream classification datasets.

The detailed comparison of these different CFG scales are presented in Table A11. Based on these results, we determined that a CFG scale of 1.5 delivers the best zero-shot classification performance on ImageNet. Consequently, we chose CFG$=1.5$ for the majority of our CLIP experiments.

H.3. Detailed experiment results for all scales

Table A12 provides detailed scaling behavior for CLIP models trained utilizing either synthetic, real, or a combination of synthetic and real images. We also present the scaling behavior comparison in detailed plots for each specific downstream dataset in Figure A16. Figure A15 shows the

Table A12. Zero-shot transfer performance on 15 downstream datasets and ImageNet. Models are trained on LAION-400M subsets with images from synthetic, real or synthetic+real. Dataset scale starts from 1M and increase exponentially. Combining synthetic images and real images can improve zero-shot classification performance under various cases, especially when data amount is limited.

Scale	Data	Food-101	CIFAR-10	CIFAR-100	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	STL-10	EuroSAT	RESISC45	GTSRB	Country211	**Average**	**ImageNet**
1M	Syn	5.2	12.8	3.3	5.9	1.7	0.9	5.5	6.7	17.8	3.5	29.4	9.0	9.7	5.4	1.2	7.9	4.1
	Real	5.2	25.4	7.6	5.0	2.1	1.0	5.4	5.4	18.0	5.0	36.4	14.7	9.3	6.6	1.0	9.9	3.8
	Syn+Real	**10.9**	**32.2**	**13.0**	**13.1**	**4.6**	**1.4**	**9.4**	**12.0**	**36.0**	**8.9**	**62.5**	**19.8**	**14.7**	**7.5**	**1.9**	**16.5**	**9.4**
2M	Syn	11.0	15.3	3.8	14.5	6.2	1.7	10.3	15.6	36.2	7.2	36.3	15.6	14.5	3.4	1.7	12.9	10.7
	Real	13.4	39.0	16.8	13	6.6	1.3	10.5	13.0	40.1	12.4	57.1	**17.0**	14.9	**6.5**	1.7	17.6	10.6
	Syn+Real	**22.2**	**59.7**	**27.0**	**23.4**	**18.2**	**2.1**	**15.4**	**24.8**	**55.7**	**13.1**	**73.9**	**22.4**	**20.6**	**4.2**	**3.0**	**25.7**	**19.8**
4M	Syn	19.6	19.7	7.1	23.0	22.4	2.1	17.0	30.1	53.4	13.9	64.4	12.8	21.1	5.3	3.1	21.0	19.8
	Real	30.8	63.8	33.4	26.2	27.7	1.8	18.8	33.9	61.7	18.9	79.2	**40.2**	21.5	10.0	3.4	31.4	21.7
	Syn+Real	**40.3**	**67.1**	**39.3**	**35.8**	**40.9**	**2.3**	**22.9**	**45.4**	**70.1**	**23.1**	**88.2**	33.5	**27.7**	**12.3**	**4.5**	**36.9**	**30.6**
8M	Syn	34.2	23.8	9.5	32.6	39.9	3.5	20.3	46.3	63.0	20.7	78.7	9.8	19.1	4.9	4.3	27.4	29.0
	Real	48.7	79.6	47.9	38.5	48.9	3.9	25.5	52.8	74.9	**31.2**	88.0	27.4	32.8	**16.7**	5.2	41.5	34.2
	Syn+Real	**54.5**	**82.9**	**53.1**	**46.3**	**57.3**	**5.1**	**29.6**	**61.4**	**78.2**	31.1	**92.5**	**29.6**	**41.1**	14.5	**6.5**	**45.6**	**40.7**
16M	Syn	44.2	32.4	11.5	41.6	51.3	4.9	27.4	58.3	72.1	24.8	83.6	16.7	29.5	4.6	5.9	33.9	37.7
	Real	62.9	85.2	58.1	49.0	60.6	**5.0**	30.4	61.9	81.5	**40.9**	93.1	**43.2**	39.4	**28.0**	7.4	49.8	43.8
	Syn+Real	**64.8**	**87.5**	**61.0**	**53.7**	**63.3**	**4.9**	**36.5**	**67.7**	**82.8**	**38.6**	**94.5**	**37.6**	**48.2**	**28.6**	**8.2**	**51.9**	**48.2**
32M	Syn	54.2	33.3	18.3	47.3	58.9	4.4	30.3	65.3	75.3	29.4	88.5	15.5	35.5	8.5	7.3	38.1	43.8
	Real	70.4	86.3	64.7	55.7	67.2	4.9	35.7	70.1	82.8	**46.0**	94.9	**43.4**	48.6	**36.6**	9.1	54.4	50.5
	Syn+Real	**71.4**	**89.4**	**65.3**	**57.9**	**69.7**	**6.5**	**41.8**	**72.9**	**83.2**	41.3	**95.2**	38.7	**55.1**	29.9	**10.2**	**55.2**	**52.9**
64M	Syn	59.7	44.1	20.9	51.5	62.7	**7.7**	37.9	71.1	79.2	35.8	92.1	15.1	39.0	12.5	8.6	42.5	48.0
	Real	75.2	**90.9**	**69.5**	59.4	71.5	7.3	42.8	75.0	87.0	**50.6**	95.7	**46.8**	51.8	**39.8**	11.2	58.3	55.1
	Syn+Real	**74.6**	90.8	67.8	**61.1**	**73.3**	6.3	**50.2**	**76.3**	**87.8**	47.6	**95.8**	45.5	**58.3**	37.5	**11.6**	**59.0**	**56.4**
128M	Syn	63.7	45.1	15.9	52.3	67.1	9.3	37.8	75.7	80.5	39.1	93.2	8.0	35.7	10.1	9.5	42.9	51.2
	Real	**81.9**	90.5	**70.9**	62.5	78.7	10.7	46.0	**85.9**	88.7	**60.4**	96.0	**48.3**	57.8	42.7	**14.2**	62.3	61.4
	Syn+Real	81.6	**91.0**	70.4	**64.0**	**79.4**	**11.9**	**52.5**	85.1	**90.2**	59.5	**97.0**	47.3	**61.1**	**45.3**	14.1	**63.4**	**62.9**
256M	Syn	68.6	46.2	21.8	54.7	70.4	10.9	42.9	80.2	81.5	44.6	95.2	20.2	39.1	12.8	10.5	46.6	54.4
	Real	**84.6**	**92.8**	**73.5**	**66.5**	**82.4**	12.3	52.7	**89.9**	91.3	**65.7**	96.9	**39.2**	64.4	**47.3**	**16.9**	**65.1**	**65.4**
	Syn+Real	**83.8**	**92.4**	**73.3**	**66.0**	**82.3**	**14.6**	**55.0**	**86.7**	**91.4**	58.6	**97.8**	**47.7**	**65.2**	42.5	**15.3**	64.8	**65.4**
371M	Syn	70.1	51.9	26.2	55.5	70.8	12.3	41.5	79.6	83.6	45.5	95.7	28.8	39.3	20.6	10.9	48.8	55.7
	Real	**85.7**	**93.9**	**75.6**	**67.5**	**83.3**	14.2	50.1	**88.8**	91.1	**67.0**	97.0	43.9	**66.6**	42.8	**17.5**	65.7	**66.8**
	Syn+Real	84.6	92.4	73.2	67.1	82.0	**17.2**	**56.8**	86.4	**91.7**	61.6	**97.3**	**52.2**	65.9	**46.7**	16.0	**66.1**	66.6

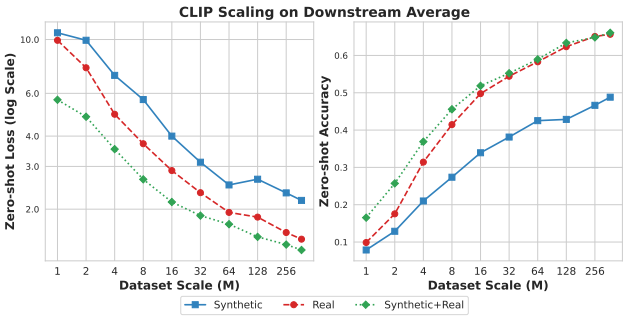


Figure A15. The average scaling behavior on zero-shot classification for CLIP models over all 15 downstream datasets. Models are trained on LAION-400M subsets with synthetic, real, or synthetic+real images.

average scaling behavior over all 15 downstream datasets. The models were trained on subsets of the LAION-400M dataset, beginning with 1 million samples and scaling up exponentially to the entire set of 371M. Our findings indicate synthetic images does not scale as good as real onees, yet integrating synthetic images with real images in the training of CLIP models can be advantageous, particularly in scenarios where the dataset size is relatively limited.

Acknowledgements

The authors would like to thank Shobhita Sundaram, Julia Chae, Sara Beery, and the VisCam team for fruitful discussions, Yuanzhen Li for helping with computation resources, and Jason Baldridge and Sergey Ioffe for guidance and check on publication policy.

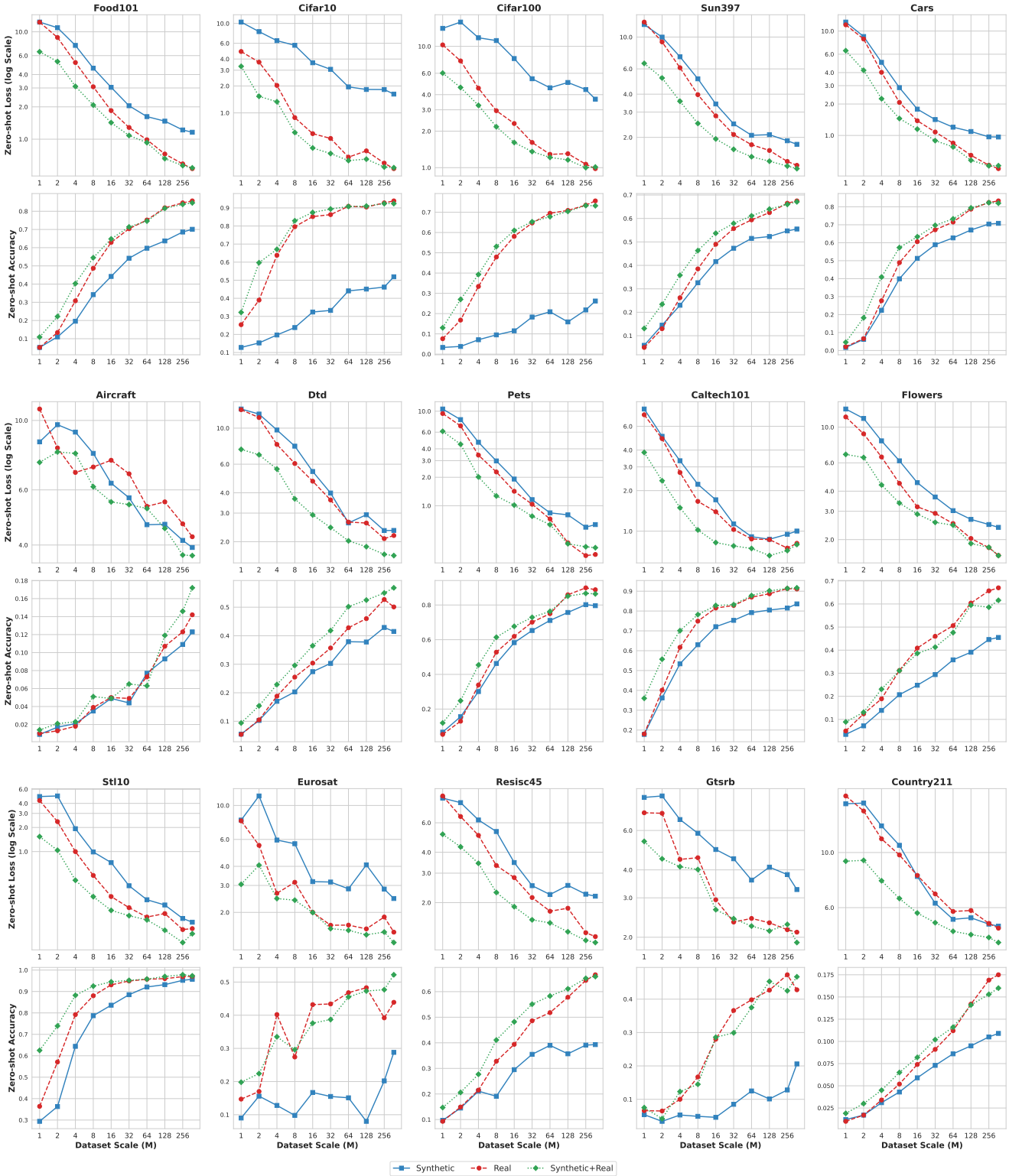


Figure A16. Detailed scaling behavior comparison between CLIP models trained on synthetic, real, or the combination of synthetic and real images, on 15 different downstream tasks. The models are evaluated under zero-shot classification.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014. 2

[2] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 13

[3] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 16

[4] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 2017. 2

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 2

[6] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, 2011. 2

[7] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR workshops*, 2020. 1

[8] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *TPAMI*, 2006. 2

[9] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. 2

[10] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021. 7

[11] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021. 7

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 7

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[14] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV workshops*, 2013. 2

[15] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2

[16] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *CVPR*, 2023. 1

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[18] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 2

[19] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *ECCV*, 2022. 1

[20] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, 2008. 2

[21] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 2

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2

[23] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. 7

[24] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022. 12

[25] Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *CVPR*, 2023. 1

[26] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *IJCNN*, 2011. 2

[27] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021. 1

[28] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 2016. 2

[29] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019. 7

[30] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 2

[31] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[32] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *NeurIPS*, 2014. 1