Prompt, Generate, then Cache:
Cascade of Foundation Models makes Strong Few-shot Learners

Renrui Zhang[*2,3], Xiangfei Hu[*2,4], Siyuan Huang[2,4], Bohao Li[5], Hanqiu Deng[2],
Hongsheng Li[3], Yu Qiao[2], Peng Gao[†1,2]

[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Science
[2]Shanghai Artificial Intelligence Laboratory   [3]The Chinese University of Hong Kong
[4]Shanghai Jiaotong University   [5]University of Chinese Academy of Sciences

{zhangrenrui, huangsiyuan, qiaoyu, gaopeng}@pjlab.org.cn,
sjtuhxf@sjtu.edu.cn, hsli@ee.cuhk.edu.hk

Abstract

*Visual recognition in low-data regimes requires deep neural networks to learn generalized representations from limited training samples. Recently, CLIP-based methods have shown promising few-shot performance benefited from the contrastive language-image pre-training. We then question, if the more diverse pre-training knowledge can be cascaded to further assist few-shot representation learning. In this paper, we propose **CaFo**, a **Ca**scade of **Fo**undation models that incorporates diverse prior knowledge of various pre-trianing paradigms for better few-shot learning. Our CaFo incorporates CLIP's language-contrastive knowledge, DINO's vision-contrastive knowledge, DALL-E's vision-generative knowledge, and GPT-3's language-generative knowledge. Specifically, CaFo works by 'Prompt, Generate, then Cache'. First, we leverage GPT-3 to produce textual inputs for prompting CLIP with rich downstream visual semantics. Then, we generate synthetic images via DALL-E to expand the few-shot training data without any manpower. At last, we introduce a learnable cache model to adaptively blend the predictions from CLIP and DINO. By such collaboration, CaFo can fully unleash the potential of different pre-training methods and unify them to perform state-of-the-art for few-shot classification. Code is available at* https://github.com/ZrrSkywalker/CaFo.

1. Introduction

Convolutional neural networks [39] and transformers [62] have attained great success on a wide range of vision tasks with abundant datasets [13]. Instead, for some

---

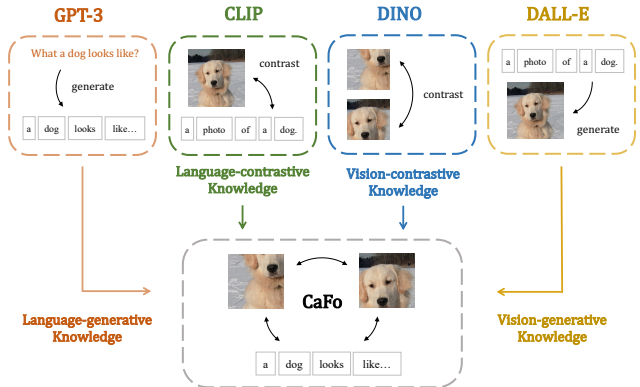* Equal contribution. † Corresponding author

Figure 1. **The Collaboration Paradigm of CaFo.** We adaptively incorporate the knowledge from four types of pre-training methods and achieve a strong few-shot learner.

data-deficient and resource-finite scenarios, few-shot learning [58, 64] also becomes a research hotspot, where the networks are constrained to learn from limited images with annotations. Many previous works have been proposed in this field to enhance model's generalization capability by meta learning [18, 65], metric learning [68], and data augmentation [26, 67]. Recently, CLIP [52] pre-trained by large-scale language-image pairs shows favorable zero-shot transfer ability for open-vocabulary visual recognition. The follow-up CoOp [76], CLIP-Adapter [20] and Tip-Adapter [70] further extend it for few-shot classification and achieve superior performance on various downstream datasets. This indicates that, even if the few-shot training data is insufficient, the large-scale pre-training has endowed the network with strong representation ability, which highly benefits the few-shot learning on downstream domains. Now that there exist various self-supervisory

paradigms besides CLIP, *could we adaptively integrate their pre-learned knowledge and collaborate them to be a better few-shot learner?*

To tackle this issue, we propose **CaFo**, a **Ca**scade of **Fo**oundation models blending the knowledge from multiple pre-training paradigms with a 'Prompt, Generate, then Cache' pipeline. As shown in Figure 1, we integrate CLIP [52], DINO [6], DALL-E [53], and GPT-3 [3] to provide four types of prior knowledge for CaFo. Therein, CLIP [52] is pre-trained to produce paired features in the embedding space for every image and its descriptive text. Guided by texts with different categorical semantics, CLIP [52] can well classify the images aided by **language-contrastive knowledge**. DINO follows contrastive self-supervised learning [6] to match the representations between two transformations of one same image, which is expert at distinguishing different images with **vision-contrastive knowledge**. Similar to CLIP [52], DALL-E [53] is also pre-trained by image-text pairs but learns to predict the encoded image tokens based on the given text tokens. Conditioned on the input text, DALL-E [53] could leverage the **vision-generative knowledge** to create high-quality synthetic images in a zero-shot manner. Pre-trained by large-scale language corpus, GPT-3 [3] takes a few hand-written templates as input, and autoregressively generates human-like texts, which contain rich **language-generative knowledge**. Therefore, the four models have distinctive pre-training goals and can provide complementary knowledge to assist the few-shot visual recognition.

In detail, we cascade them by three steps.: **1) Prompt**. We adopt GPT-3 [3] to produce textual prompts for CLIP based on a few hand-written templates. These prompts with richer language knowledge are fed into CLIP's textual encoder. **2) Generate**. We adopt DALL-E [53] to generate additional training images for different categories based on the domain-specific texts, which enlarges the few-shot training data, but costs no extra manpower for collection and annotation. **3) Cache**. We utilize a cache model to adaptively incorporate the predictions from both CLIP [52] and DINO [6]. Referring to Tip-Adapter [70], we build the cache model with two kinds of keys respectively for the two pre-trained models. Regarding zero-shot CLIP as the distribution baseline, we adaptively ensemble the predictions of two cached keys as the final output. By only fine-tuning the lightweight cache model via expanded training data, CaFo can learn to fuse diverse prior knowledge and leverage their complementary characteristics for better few-shot visual recognition.

Our main contributions are summarized as follows:

- We propose CaFo to incorporate the prior knowledge learned from various pre-training paradigms for better few-shot learning.

- By collaborating CLIP, DINO, GPT-3 and DALL-E, CaFo utilizes more semantic prompts, enriches the limited few-shot training data, and adaptively ensembles diverse predictions via the cache model.

- We conduct thorough experiments on 11 datasets for few-shot classification, where CaFo achieves *state-of-the-art* without using extra annotated data.

2. Related Work

Pre-training of Vision Models. With the breakthroughs in deep learning models [16, 30, 43], most modern vision models are based on the paradigm of pre-training on ImageNet [13] and fine-tuning on downstream tasks [29]. Pre-trained models have shown promising adaptability for various downstream tasks, such as object detection [41], semantic segmentation [7], and 3D recognition [24, 71, 74]. To improve the representation capability by overcoming the constraints of annotation, self-supervised pre-training has attracted wide attention using large-scale unlabeled datasets [37]. Self-supervised learning is initialized by pretext tasks, such as image restoration from corruption [27, 50, 63], pseudo labels [15, 48] and clustering [4]. Recently, contrast learning, which learns representations by contrasting positive pairs against negative pairs, has gotten well studied for diverse visual representation learning [6,8,9,23,28,61]. Besides, language-supervised visual pre-training emerges as a novel paradigm closer to natural visual understanding [1, 46, 54, 57], among which CLIP [52] obtains powerful zero-shot transferability by contrastive pre-training on image-text pairs from the Internet. In addition, vision-language pre-training can also promote the zero-shot image generation from text. Open generative models, such as DALL-E [53] and CogView [14] pre-trained on large-scale image-text pairs are able to generate images with diverse contents by given texts. In this paper, CaFo cascade three visual pre-training models, CLIP, DINO, and DALL-E, which contributes to better few-shot learning capacity.

Language-assisted Vision Models. As different form of data, linguistic knowledge normally contains complementary knowledge to images. For vision-language models, several works [3,21,52] have showed the format of prompts would highly affect the accuracy on vision tasks. Thus, prompt engineering is worth putting in great effort. Some efforts [34,55,73,77] utilize learnable textual inputs and optimize them during training. Other works [45, 51] propose to leverage linguistic knowledge pre-trained from large language models to generate prompts for each visual category, which enhances vision-language models without any additional training or labeling. Our CaFo refers to CuPL [51] to produce semantic-rich texts to prompt CLIP for better text-image alignment.

Few-shot Adaptation. Few-shot learning highly relies on the transferability of the trained neural networks. From the perspective of distance measurement, some metric learning methods learn a metric space by computing the distances from the instances to novel categories [58, 60, 64]. Also, meta-learning is proposed to improve the few-shot adaptation ability of the models by finding a set of initialized parameters that can rapidly adapt to novel domains [10, 19, 35, 40]. More recently, with the vision-language pre-training model CLIP [52] exhibiting strong zero-shot adaptation performance, several efforts have started to find efficient strategies to adapt it to downstream few-shot datasets. Following the recent research trend of NLP in prompt learning, CoOp [76] is proposed as a prompt tuning adaptation method by optimizing a set of learnable prompt tokens. Subsequently, to address the harm of generalization ability brought by CoOp [76], that is, the recognition of unknown categories becomes poor, CoCoOp [78] and VT-CLIP [73] propose to train a meta-network to generate image tokens as conditional inputs for the textual vectors. Referring to adapters [33] in NLP, CLIP-Adapter [20] is introduced to fine-tune CLIP by applying lightweight residual-style bottleneck layers as the adapter. Tip-Adapter [70] is then proposed as a training-free adaption method with a constructed key-value cache model. It can also be regarded as a better initialization of CLIP-Adapter with much faster convergence when fine-tuning. CALIP [25] proposes a parameter-free attention to enhance CLIP in a zero-shot manner, and its parametric solution further attains higher few-shot accuracy. Besides, many follow-up works [34, 42, 69, 72, 75, 79] have also been proposed for further adapting CLIP to various vision tasks. Different from all existing methods, we integrate other powerful pre-training paradigms with CLIP and collaborate them with customized pipelines.

3. Cascade of Foundation Models

In this section, we first briefly revisit four types of pre-training paradigms in CaFo. Then, we specifically introduce how we cascade them by 'Prompt, Generate, then Cache'.

3.1. Different Pre-training Paradigms

Contrastive Vision-Language Pre-training. The series [8] of contrastive learning between vision and language learn to map the two modalities into the same embedding space via a contrastive loss. Driven by web-scale datasets, e.g., 400 million for CLIP [52] and 1.8 billion for ALIGN [36], the basic pre-training target is to minimize the embedding distances of images and their textual descriptions, while maximize those unpaired ones. By the cross-modal alignment, we can discriminate images of different categorizes by the texts with different semantics. We denote such learned prior as language-contrastive knowledge
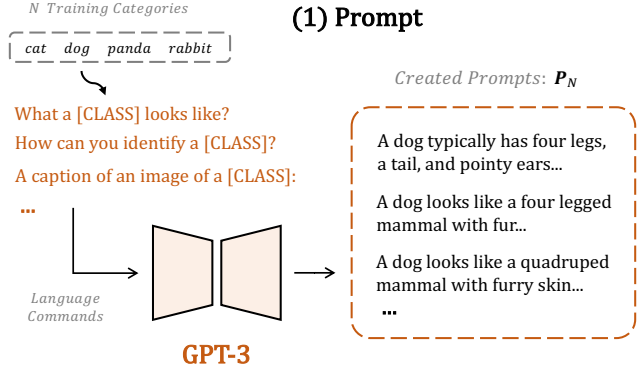

Figure 2. **Prompt with GPT-3 [3]**. As the first step in CaFo, we utilize the pre-trained GPT-3 to produce prompts with rich linguist semantics for CLIP's textual encoder.

and adopt CLIP as the representative model for such pre-training method.

Contrastive Vision Pre-training. As the traditional self-supervised learning methods, vision-contrastive models [8] focus on the discrimination between different images. Normally, the positive pairs to be drawn close are two transformations of the same image, while the optimization of negative pairs [22] is optional, which can be replaced by a momentum encoder [28] or cluster assignments [5]. Recent works reveal that we can learn self-supervised features without negative pairs between images [6, 23]. Given the strong linear classification capacity, the pre-trained DINO [6] is adopted here to provide vision-contrastive knowledge for collaboration.

Generative Language Pre-training. With 175 billion parameters, the large-scale pre-trained GPT-3 [3] is powerful to produce human-like texts with diverse contents and incredible quality. Taking as input a few designed language commands, GPT-3 is able to output prompts with rich linguistic semantics for vision-language models. CLIP utilizes handcrafted templates as prompts, e.g., "a photo of a [CLASS]", which however lacks sufficient textual semantics to align with input images. We thus leverage GPT-3 to produce CLIP's prompts to better align with visual information from images.

Generative Vision-Language Pre-training. Learned from millions of image-caption pairs, the DALL-E series can generate language-conditioned images in a zero-shot manner. They are pre-trained to autoregressively predict the encoded image tokens from the textual tokens of the captions. With such language-generative knowledge, the pre-trained DALL-E can be viewed as a free lunch to enlarge the training data without any manpower. Considering publicity, we select DALL-E-mini [12] as the representative among DALL-E models.
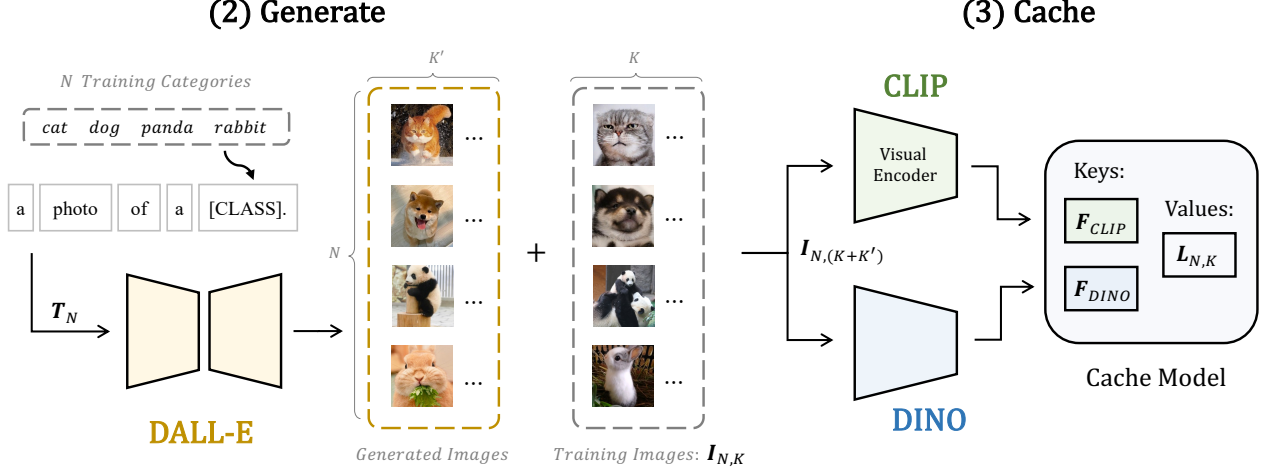
Figure 3. **Generate via DALL-E [53], then Cache by CLIP [52] and DINO [6].** We adopt DALL-E to generate synthetic images to expand the limited few-shot training samples. Then, we construct the cache model with two kinds of keys to adaptively fuse the knowledge from CLIP and DINO.

3.2. Prompt, Generate, then Cache

To cascade different pre-training paradigms, we introduce CaFo with a pipeline of 'Prompt, Generate, then Cache', which respectively unleashes the powers of different self-supervised knowledge.

Prompt with GPT-3. Under the N-way K-shot settings, we have the few-shot training images $I_{N,K}$ with labels $L_{N,K}$ that contain K samples for each N categories. As shown in Figure 2, for N categories, we adopt a unified series of templates as the language command for GPT-3 [3], e.g., "What a [CLASS] looks like?", "How can you identify a [CLASS]?", and "A caption of an image of a [CLASS]:". We denote the created prompts for N categories as P_N, formulated as

$$P_N = \text{GPT-3}(\text{Commands}). \quad (1)$$

Then, we adopt P_N as the input of CLIP's textual encoder. Further, for some downstream data with specialized categories, we can customize the language commands for producing prompts with more domain-specific semantics. For example, in OxfordPets [49] dataset of pet images, we adopt the input of GPT-3 as "This is a pet bulldog, it has thin neck, short face, floppy ears. It's coat is short, straight, and in brindle color. This is a pet [CLASS],". Based on that, GPT-3 continues to describe more details of the [CLASS] pet.

Generate via DALL-E Via the zero-shot DALL-E [12], we generate synthesis images to enrich our limited training images $I_{N,K}$, as shown in Figure 3 (1). For different categories, we adopt a simple template, e.g., "a photo of

a [CLASS].". After the generation, we utilize CLIP to filter the top-K' best-quality images as the newly-expanded training samples for each category. Then, we obtain the N-category $(K + K')$-sample training images, formulated as

$$I_{N,(K+K')} = \{\text{DALL-E}(T_N), \ I_{N,K}\},, \quad (2)$$

where T_N denotes the N-category textual inputs. We keep K' comparable with K to ensure the synthesis quality and also preserve the low-data regimes. By the pre-trained language-generative knowledge, the data expansion is totally zero-shot, which requires no manpower to collect or annotate the data, and alleviates the data deficiency issue inherently for few-shot learning.

Cache by CLIP and DINO. We construct a key-value cache model for adaptive knowledge ensemble. Different from Tip-Adapter [70] only adapting CLIP, our cache model contains the pre-learned knowledge from both CLIP and DINO by caching two kinds of keys. Specifically in Figure 4 (2), we first utilize CLIP and DINO to independently extract visual features of the few-shot training images, formulated as

$$F_{\text{CLIP}} = \text{CLIP}_{vis}(I_{N,(K+K')}); \quad (3)$$
$$F_{\text{DINO}} = \text{DINO}(I_{N,(K+K')}), \quad (4)$$

where CLIP_{vis} denotes the CLIP's visual encoder and $F_{\text{CLIP}}, F_{\text{DINO}} \in \mathbb{R}^{N(K+K') \times C}$. Besides the two keys, we convert the few-shot training labels into one-hot encodings $L_{onehot} \in \mathbb{R}^{N(K+K') \times N}$, and regard them as the same values for both keys. During training, we follow Tip-Adapter that only enables the cached keys in the adapter to be learnable and keeps the pre-trained models frozen.
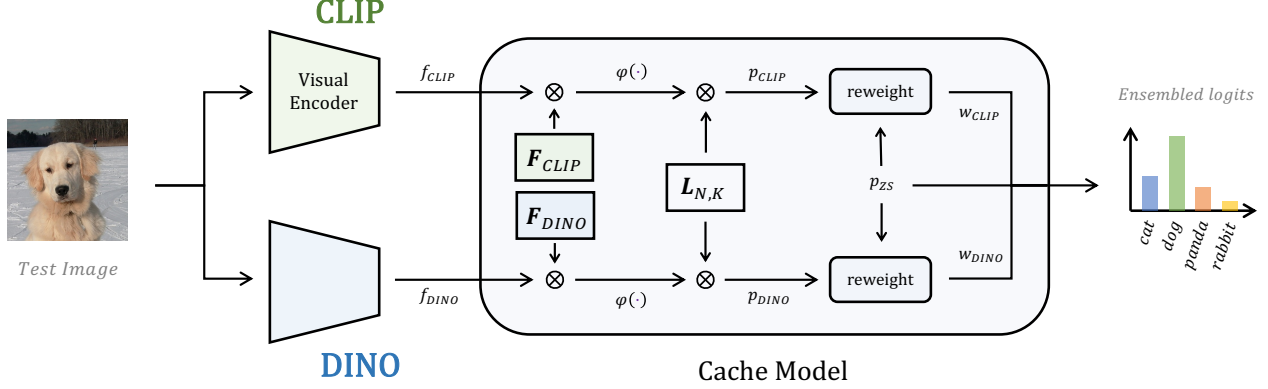
Figure 4. **Adaptive Inference with Cache Model.** We regard the test image as a query and retrieves CLIP and DINO's knowledge from the corresponding two keys in the cache model. Then, we calculate the distribution similarities between different classification logits for adaptive ensemble.

3.3. Adaptive Inference

For a test image in Figure 4, we first extract its two visual features $f_{\text{CLIP}}, f_{\text{DINO}} \in \mathbb{R}^{1 \times C}$ and regard them as queries to retrieve diverse knowledge from the cache model. Then, we could acquire three predicted classification logits $p_{\text{ZS}}, p_{\text{CLIP}}, p_{\text{DINO}} \in \mathbb{R}^{1 \times N}$, which are respectively from CLIP's zero-shot alignment and the two keys of cache model. We formulate them as

$$p_{\text{ZS}} = f_{\text{CLIP}} \, \text{CLIP}_{tex}(P_N)^T; \tag{5}$$

$$p_{\text{CLIP}} = \varphi(f_{\text{CLIP}} F_{\text{CLIP}}^T) \, L_{onehot}; \tag{6}$$

$$p_{\text{DINO}} = \varphi(f_{\text{DINO}} F_{\text{DINO}}^T) \, L_{onehot}, \tag{7}$$

where CLIP_{tex} represents CLIP's textual encoder, P_N denotes GPT-3's created prompts, and $f_{\text{CLIP}} F_{\text{CLIP}}^T$ denotes the query-key affinity matrix of the CLIP's keys, analogous to DINO's. $\varphi(x) = \exp(-\beta \cdot (1 - x))$ serves as a non-linear modulator to control the sharpness of affinity matrix.

As the language-contrastive p_{ZS} is pre-trained by 400 million data and can perform strong zero-shot transfer ability, we regard p_{ZS} as the prediction baseline and calculate the weights of $p_{\text{CLIP}}, p_{\text{DINO}}$ for ensemble based on their distribution similarity with p_{ZS}. By this, we can suppress some obviously false category possibilities in $p_{\text{CLIP}}, p_{\text{DINO}}$ and also amplify the moderately correct ones during ensemble. Firstly, we respectively normalize the scales of three classification logits into -1~1 by their each mean and standard deviation. We then calculate the distribution similarities as the ensemble weights for the two logits of the cache as

$$w_{\text{CLIP}} = p_{\text{CLIP}} \, p_{\text{ZS}}^T; \quad w_{\text{DINO}} = p_{\text{DINO}} \, p_{\text{ZS}}^T. \tag{8}$$

Finally, we adopt the softmax function to normalize the weights and obtain the final ensemble logits as

$$p_{en} = p_{\text{ZS}} + \sum_i p_i \cdot \text{softmax}(w_i), \tag{9}$$

where $i \in \{\text{CLIP}, \text{DINO}\}$. By such similarity-based ensemble, p_{en} can adaptively fuse the prior knowledge learned by CLIP and DINO's pre-training and achieve stronger few-shot image classification.

4. Experiments

4.1. Settings

Datasets. We conduct few-shot experiments on 11 publicly available datasets: ImageNet [13], Standford-Cars [38], UCF101 [59], Caltech101 [17], Flowers102 [47], SUN397 [66], DTD [11], EuroSAT [31], FGVCAircraft [44], OxfordPets [49], and Food101 [2]. We follow Tip-Adapter [70] to train CaFo with 1, 2, 4, 8, 16 shots and test on the full test set. *As we adopt DALL-E to generate training images in a zero-shot manner, we can train CaFo only by the generated images and report its zero-shot performance without few-shot training set.*

Implementation. Our CaFo integrates the knowledge from pre-trained CLIP [52], DINO [6], DALL-E [53], and GPT-3 [3]. For CLIP, we utilize ResNet-50 [30] as the visual encoder and its aligned transformer as the textual encoder. To align with the visual representation from CLIP, we also adopt DINO pre-trained upon ResNet-50. For DALL-E, we adopt different domain-specific textual templates as the input for different datasets, which correspond to the original textual prompts for CLIP's textual encoder. For GPT-3, we adopt five simple templates as the language commands shared by different categories. Each command outputs ten prompts, which obtains fifty prompts in total. For each category, we simply ensemble the features of different prompts following CuPL [3]. During training, we only set the two kinds of keys in cache model to be learnable and utilize the data augmentation following Tip-Adapter-F. We train CaFo using batch size 64 only for 20 epochs, and adopt AdamW optimizer with the initial learning rate
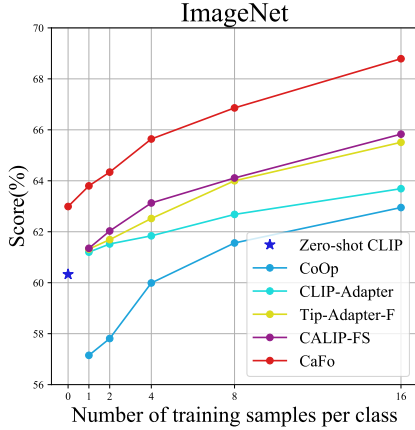
Figure 5. **Performance (%) Comparison on ImageNet.** We compare CaFo with other methods for different few-shot settings.

Models	Epochs	Time	Accuracy	Gain
Zero-shot CLIP	0	0	60.33	-
Zero-shot CALIP	0	0	60.57	-
Linear-probe CLIP	-	13min	56.13	-4.20
CoOp	200	14h 40min	62.95	+2.62
CLIP-Adapter	200	50min	63.59	+3.26
Tip-Adapter-F	**20**	**5min**	65.51	+5.18
CALIP-FS	200	1h	65.81	+5.48
CaFo	**20**	10min	**68.79**	**+8.46**

Table 1. **Efficiency Comparison on ImageNet.** We test the training time with a single A100 GPU under 16-shot setting.

Shot	0	1	2	4	8	16
Zero-shot CLIP	60.33	-	-	-	-	-
Zero-shot CALAP	60.57	-	-	-	-	-
Linear-probe CLIP	-	22.17	31.90	41.20	49.52	56.13
CoOp	-	57.15	57.81	59.99	61.56	62.95
CLIP-Adapter	-	61.20	61.52	61.84	62.68	63.59
VT-CLIP	-	60.53	61.29	62.02	62.81	63.92
Tip-Adapter-F	-	61.32	61.69	62.52	64.00	65.51
CALIP-FS	-	61.35	62.03	63.13	64.11	65.81
CaFo	**62.99**	**63.80**	**64.34**	**65.64**	**66.86**	**68.79**

Table 2. **Quantative Performance (%) Comparison on ImageNet.** For zero-shot performance, CaFo is trained with images generated by DALL-E without any few-shot data.

Datasets	Source	Target	
	ImageNet	-V2	-Sketch
Zero-shot CLIP	60.33	53.27	35.44
Zero-shot CALIP	60.57	53.70	35.61
CoOp	62.95	54.58	31.04
CLIP-Adapter	63.59	55.69	35.68
CALIP-FS	65.81	55.98	35.37
Tip-Adapter-F	65.51	57.11	36.00
CaFo	**68.79**	**57.99**	**39.43**

Table 3. **Distribution Shift (%) Comparison.** We train the models on "Source" dataset and test on "Target" datasets.

0.0001 with a cosine scheduler. Note that, we tune the hyperparameters in CaFo by the official validation sets.

4.2. Performance

On ImageNet. We compare CaFo with other CLIP-based adaption methods on the most representative ImageNet [13]: CALIP [25], Linear-probe CLIP [52], CoOp [76], CLIP-Adapter [20], Tip-Adapter-F [70], and CALIP-FS [25]. All these methods are based on the pre-trained CLIP [52] with ResNet-50 visual encoders. As reported in Figure 5 and Table 2, CaFo surpasses all existing methods for different shot settings. Remarkably, CaFo with 1 shot even outperforms the 8-shot Linear-probe CLIP and CoOp, and CaFo with 8 shots is better than all methods with 16 shots. For zero-shot learning, CaFo sigiificantly surpasses CLIP and CALIP, demonstrating the importance of DALL-E's generation. In Table 1, we present the efficiency of CaFo concerning training epochs and time. Our CaFo achieves the best performance-efficiency trade-off with 68.79% accuracy and only 10 minutes training.

On Other Datasets. To further assess the robustness in different scenarios, we test CaFo on extra 10 datasets in Figure 6. For different semantic domains including real-world scenes, detailed textures, and satellite-captured landscapes, CaFo consistently shows leading performance and indicates excellent robustness via the collaboration of diverse knowledge. Notably, on some datasets, e.g., Caltech101 and OxfordPets, the zero-shot CaFo perform even comparably to other methods with 4 shots, demonstrating the effectiveness of zero-shot DALL-E for few-shot data expansion.

Distribution Shift. We further evaluate the robustness of CaFo to distribution shift by training on "Source" dataset and testing on "Target" datasets. In Table 3, we select the "Source" as ImageNet and the "Target" as ImageNet-V2 [56] and ImageNet-Sketch [32]. As we can utilize some prior knowledge of the target domain for GPT-3 and DALL-E for prompting and generation, CaFo achieves the best out-of-distribution performance on the two "Target" datasets, surpassing the second-best Tip-Adapter-F by +3.28%, +0.88%, and +3.43%, respectively.
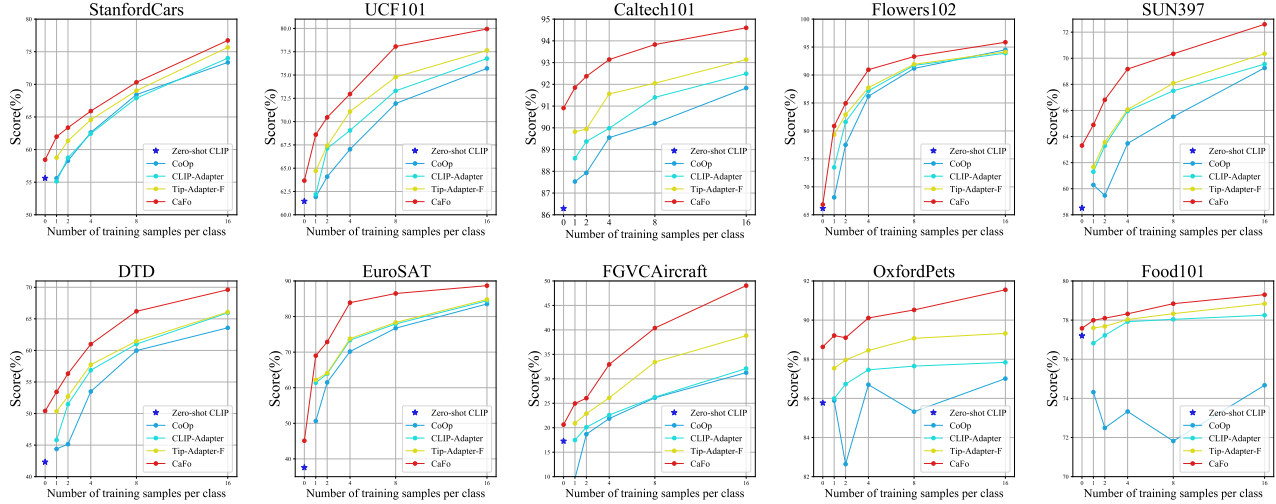
Figure 6. **Performance (%) Comparison on 10 Datasets.** Our method shows *state-of-the-art* performance for all few-shot settings on different datasets, which indicates superior generalization capacity.

Pre-trained Models				Shot		
CLIP	DINO	DALL-E	GPT-3	1	4	16
✓				61.32	62.52	65.51
	✓			34.14	40.47	53.27
✓	✓			61.39	63.96	68.08
✓	✓	✓		63.24	65.23	68.42
✓	✓		✓	62.32	64.64	68.51
✓	✓	✓	✓	**63.80**	**65.64**	**68.79**

Table 4. **Ablation Study (%) of Cascaded Models.** We ablate different pre-trained models on ImageNet with 1, 4, and 16 shots.

Method	Shot				
	1	2	4	8	16
CLIP	61.36	61.78	62.83	64.04	65.53
DINO	34.13	34.44	41.12	45.01	53.63
Average	60.70	60.72	60.99	61.47	61.97
Maximum	61.64	62.45	62.95	63.60	64.97
p_{CLIP} Base.	62.36	63.22	64.11	65.50	67.40
p_{DINO} Base.	62.61	63.39	64.31	65.83	67.73
p_{ZS} Base.	**63.80**	**64.34**	**65.64**	**66.86**	**68.79**

Table 5. **Ablation Study (%) of Adaptive Inference.** We conduct different ensemble methods of cache model on ImageNet.

4.3. Ablation study

Cascaded Models. In Table 4, we explore how each pre-trained model contributes to the collaboration on different shots of ImageNet. Therein, "CLIP" denotes the zero-shot CLIP with cache model containing only CLIP's keys, and "DINO" denotes only the cache model with DINO's keys. As shown in the first three rows, the CLIP's language-contrastive knowledge performs stronger than DINO's vision-contrastive knowledge, which might benefit from millions of pre-training data. Their adaptive ensemble by cache model can bring larger improvement when the shot number increases. For the next two rows, DALL-E and GPT-3 can independently boost both CLIP and DINO for nearly all shots with the prompts and generated synthetic images. The last row represents our final solution, CaFo that incorporates all three pre-trained models with the best performance for all shots.

Generated Number via DALL-E. We utilize DALL-E to generate synthetic images as the expanded few-shot training data. In Table 6, we explore the best synthetic number K' for each cat-

egory of different shots on ImageNet. We observe that the larger K' does not lead to better few-shot performance. As we adopt pre-trained CLIP to select the top-K' generated images, which are scored by the similarities between CLIP-encoded images and category texts, the larger K' would contain more low-quality images and adversely affect the cache model. Furthermore, the amount of expanded data is comparable to the original K shots and thus preserves the characteristic of few-shot learning.

Adaptive Inference. In Table 5, we ablate different ensemble methods of CLIP and DINO's predictions during inference on ImageNet. The first two rows represent the cache model with one type of keys respectively for two pre-trained models without ensemble. Then, we adopt average and maximum pooling between the two predictions and ensemble the result with p_{ZS}. However, such naive integration without adaptive weights causes accuracy degradation. In the last three rows, we calculate the distribution similarities for adaptive ensemble and respectively select the three logits as the baseline. As shown, using p_{ZS} as the distribution base-
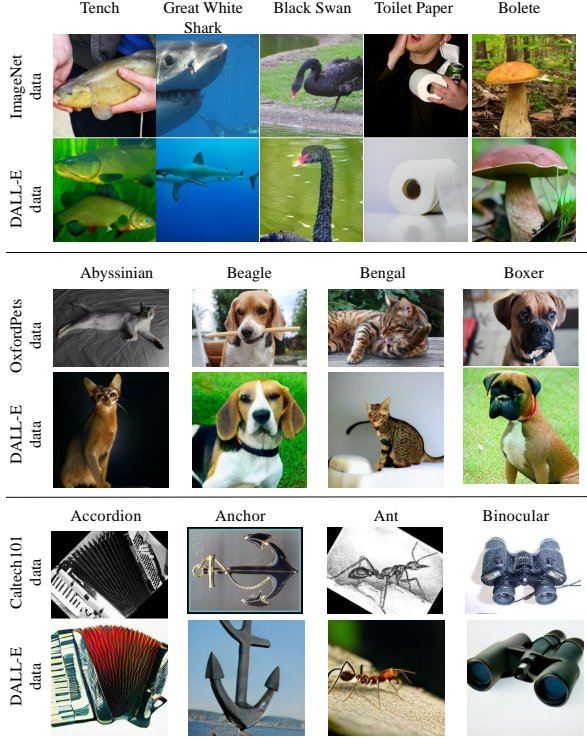
Figure 7. **Visualizations of DALL-E's Generated Images.** Examples are from ImageNet, OxfordPets and Caltech101 datasets.

Figure 8. **Visualization of GPT-3's Prompts for CLIP.** The example shown is from the ImageNet dataset.

DALL-E	1	2	4	8	16
1	63.29	64.06	65.11	66.48	68.64
2	63.66	**64.34**	65.37	**66.86**	**68.79**
4	63.71	64.33	65.35	66.75	68.61
8	**63.80**	64.26	**65.64**	66.68	68.76
16	63.68	64.16	65.40	66.57	68.41

Table 6. **Ablation Study (%) of Generated Number via DALL-E.** We compare different shot numbers on ImageNet.

Models	RN50	RN101	ViT-B/32	ViT-B/16
Zero-shot CLIP	60.33	62.53	63.80	68.73
CoOp	62.95	66.60	66.85	71.92
CLIP-Adapter	63.59	65.39	66.19	71.13
Tip-Adapter-F	65.51	68.56	68.65	73.69
CaFo	**68.79**	**70.86**	**70.82**	**74.48**

Table 7. **Ablation Study (%) of CLIP's Visual Encoders.** We experiment different visual backbones on the 16-shot ImageNet.

line performs the best, since p_{ZS} itself shows strong transfer ability and can effectively suppress the wrong predictions of other logits.

CLIP's Visual Encoders. We conduct CaFo with different CLIP's visual encoders for comparison with other methods. As shown in Table 7, CaFo consistently achieves leading performance with different visual backbones, indicating our generalizability to network architectures.

4.4. Visualization

DALL-E's Generated Images. In Figure 7, we visualize the synthetic images generated by DALL-E on ImageNet [13], OxfordPets [49] and Caltech101 [17]. As shown, benefited from the vision-generative knowledge, the generated images can well highlight the downstream semantics of target category and effectively expand the few-shot training set in low-data regimes.

GPT-3's Prompts for CLIP. In Figure 8, We present a rectified example in ImageNet [13] aided by GPT-3's prompts in CaFo. As shown, prompting by GPT-3 (Left) produces more semantic texts compared to CLIP's handcrafted templates(Right), and better depicts the visual appearances in the image, which predicts the correct category of goldfish.

5. Conclusion

We propose CaFo, a cascade of foundation models that comprehends diverse knowledge from different pre-training and follows the 'Prompt, Generate, then Cache' pipeline. We first incorporate the generative language model, GPT-3, for prompting CLIP with more semantic texts, and adopt DALL-E to expand the few-shot training data. Then, we adaptively fuse the vision-contrastive DINO with CLIP via a unified cache model. By collaboration, CaFo achieves *state-of-the-art* performance for few-shot learning

on 11 datasets. Although CaFo has unified four types of pre-training, our future direction will focus on integrating more existing pre-trained knowledge, such as the masked-generated knowledge of MAE [27] and the 3D-reconstruction knowledge of I2P-MAE [74].

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 2

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014. 5

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2, 3, 4, 5

[4] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 3

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, October 2021. 2, 3, 4, 5

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018. 2

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. 2, 3

[9] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. 2020. 2

[10] Yinbo Chen, Zhuang Liu, Huijuan Xu, Trevor Darrell, and Xiaolong Wang. Meta-baseline: Exploring simple meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9062–9071, October 2021. 3

[11] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 5

[12] Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phuc Le Khac, Luke Melas, and Ritobrata Ghosh. Dall·e mini. 7 2021. 3, 4

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 2, 5, 6, 8

[14] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 19822–19835. Curran Associates, Inc., 2021. 2

[15] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 2

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 2020. 2

[17] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 5, 8, 12

[18] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 1

[19] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR, 06–11 Aug 2017. 3

[20] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. 2021. 1, 3, 6

[21] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020. 2

[22] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach

to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. 3

[23] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. 2, 3

[24] Ziyu Guo, Xianzhi Li, and Pheng Ann Heng. Joint-mae: 2d-3d joint masked autoencoders for 3d point cloud pre-training. *arXiv preprint arXiv:2302.14007*, 2023. 2

[25] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. *arXiv preprint arXiv:2209.14169*, 2022. 3, 6

[26] Bharath Hariharan and Ross B. Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *IEEE ICCV*, pages 3037–3046, 2017. 1

[27] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, June 2022. 2, 9

[28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 3

[29] Kaiming He, Ross Girshick, and Piotr Dollar. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2, 5

[31] 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*, 12(7):2217–2226, 2019. 5, 12

[32] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 6

[33] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 3

[34] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022. 2, 3

[35] Muhammad Abdullah Jamal and Guo-Jun Qi. Task agnostic meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[36] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 3

[37] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4037–4058, 2021. 2

[38] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 5

[39] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 1

[40] Bohao Li, Boyu Yang, Chang Liu, Feng Liu, Rongrong Ji, and Qixiang Ye. Beyond max-margin: Class margin equilibrium for few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7363–7372, 2021. 3

[41] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[42] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *European Conference on Computer Vision*, pages 388–404. Springer, 2022. 3

[43] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October 2021. 2

[44] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5

[45] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022. 2

[46] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[47] 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*, pages 722–729. IEEE, 2008. 5

[48] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 69–84, Cham, 2016. Springer International Publishing. 2

[49] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 4, 5, 8

[50] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[51] Sarah Pratt, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. *arXiv preprint arXiv:2209.03320*, 2022. 2

[52] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 4, 5, 6

[53] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 18–24 Jul 2021. 2, 4, 5

[54] Cornelius Rampf, Barbara Villone, and Uriel Frisch. How smooth are particle trajectories in a cdm universe? 2015. 2

[55] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 2

[56] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 6

[57] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. 2

[58] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 1, 3

[59] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5

[60] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[61] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. 2018. 2

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 1

[63] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 1096–1103, New York, NY, USA, 2008. Association for Computing Machinery. 2

[64] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 1, 3

[65] Yu-Xiong Wang and Martial Hebert. Learning to learn: Model regression networks for easy small sample learning. In *ECCV*, pages 616–634, 2016. 1

[66] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 5

[67] Shuo Yang, Lu Liu, and Min Xu. Free lunch for few-shot learning: Distribution calibration. *arXiv preprint arXiv:2101.06395*, 2021. 1

[68] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12203–12213, 2020. 1

[69] Renrui Zhang, Hanqiu Deng, Bohao Li, Wei Zhang, Hao Dong, Hongsheng Li, Peng Gao, and Yu Qiao. Collaboration of pre-trained models makes better few-shot learner. *arXiv preprint arXiv:2209.12255*, 2022. 3

[70] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 1, 2, 3, 4, 5, 6

[71] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training. *arXiv preprint arXiv:2205.14401*, 2022. 2

[72] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Pro-*

11

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8552–8562, 2022. 3

[73] Renrui Zhang, Longtian Qiu, Wei Zhang, and Ziyao Zeng. Vt-clip: Enhancing vision-language models with visual-guided texts. *arXiv preprint arXiv:2112.02399*, 2021. 2, 3

[74] Renrui Zhang, Liuhui Wang, Yu Qiao, Peng Gao, and Hongsheng Li. Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. *arXiv preprint arXiv:2212.06785*, 2022. 2, 9

[75] Renrui Zhang, Ziyao Zeng, Ziyu Guo, and Yafeng Li. Can language understand depth? In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6868–6874, 2022. 3

[76] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. 2021. 1, 3, 6

[77] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021. 2

[78] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16816–16825, June 2022. 3

[79] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyao Zeng, Shanghang Zhang, and Peng Gao. Pointclip v2: Adapting clip for powerful 3d open-world learning. *arXiv preprint arXiv:2211.11682*, 2022. 3

A. Additional Performance Comparison

In Figure 9, we compare the performance of CaFo without DALL-E's generated images or GPT-3's created prompts on 10 datasets, which still consistently outperform the second-best Tip-Adapter-F.

B. Additional Ablation Study

Zero-shot CaFo. As we leverage the pre-trained DALL-E to generate the supplementary few-shot training set in a zero-shot manner, our CaFo can be evaluated under zero-shot settings the same as CLIP, for which none of the human-annotated training images is given. In Table 9, we report the best generated image number K' of DALL-E for zero-shot CaFo. The number "0" denotes Zero-shot CLIP. For different datasets, the best number varies ranging from 1∼16, and the larger number normally cannot get the better result, probably due to the low-quality synthetic images. On Caltech101 [17] and EuroSAT [31], zero-shot CaFo largely surpasses CLIP by +4.62% and +7.54%, indicating our superiority under zero-shot settings.

Hyperparameter β. In Formula 5 and 6, we utilize a nonlinear modulator $\varphi(x) = \exp(-\beta \cdot (1-x))$ for the affinity matrix of CLIP and DINO in the cache model, where β controls the matrix sharpness. In Table 8, we experiment CaFo with different β on 16-shot ImageNet and observe 0.6 performs the best.

Sharpness β	0.4	0.5	0.6	0.7	0.8	1.0
CaFo	68.66	68.75	**68.79**	68.73	68.69	68.66

Table 8. **Ablation Study (%) of Hyperparameter β.**

C. Additional Visualization

GPT-3's Prompts for CLIP. In Figure 12 and 13, we show more visualization of the prompts produced by GPT-3 and how they assist our CaFo to rectify false predictions of the original CLIP's templates.

DALL-E's Generated Images. In Figure 14, we visualize more synthetic images generated by DALL-E on different datasets. Benefited from the pre-trained DALL-E, the generated images can well highlight the semantics of target category and effectively expand the few-shot training set in low-data regimes.

t-SNE. We present the t-SNE visualization of our CaFo and the second-best Tip-Adapter-F in Figure 10. CaFo shows more contrastive distribution of category clusters and well mitigates some aliasing between similar classes.

Learning Curves. In Figure 11, we visualize the 20-epoch learning curves of test accuracy on 16-shot ImageNet. Compared to the single CLIP, collaborating with DALL-E, DINO and GPT-3 significantly improves the convergence speed and classification accuracy on test set.
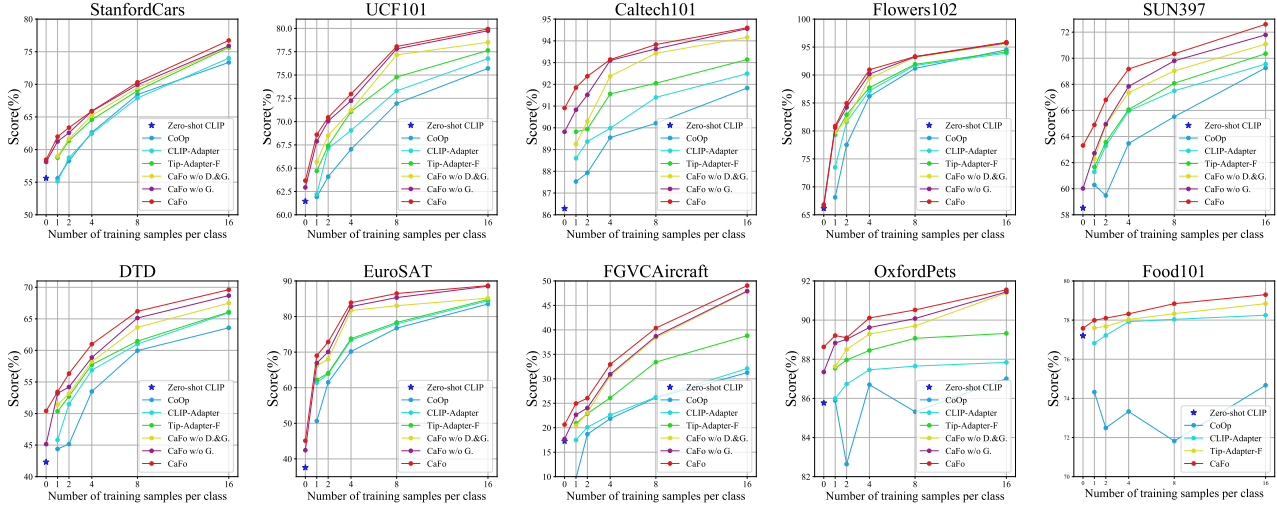
Figure 9. **Performance (%) Comparison on 10 Datasets.** Our method shows *state-of-the-art* performance for all few-shot settings on different datasets. 'CaFo w/o D.&G.' denotes CaFo without DALL-E's generated images and GPT3's created prompts.

DALL-E	ImageNet	Caltech101	Flower102	Food101	DTD	EuroSAT	OxfordPets	SUN397	StanfordCars	UCF101	FGVCAircraft
0	60.33	86.29	66.14	77.20	50.30	37.56	85.77	58.52	55.61	61.46	17.28
1	62.5	89.78	65.65	77.52	50.12	37.46	87.33	63.08	57.33	63.05	20.46
2	62.69	90.26	**66.83**	77.50	50.00	41.73	87.49	63.02	57.63	62.44	20.31
4	62.81	89.98	66.50	**77.58**	**50.41**	43.2	87.71	**63.31**	57.46	63.12	20.64
8	**62.99**	90.67	66.83	77.56	50.12	**45.10**	**88.63**	63.26	58.03	62.83	20.49
16	62.74	**90.91**	66.54	77.53	50.24	42.73	87.49	63.16	**58.45**	**63.67**	**21.06**

Table 9. **Ablation Study (%) of Zero-shot CaFo via DALL-E on Different Datasets.** We leverage DALL-E to generate different numbers of synthetic images for zero-shot recognition.
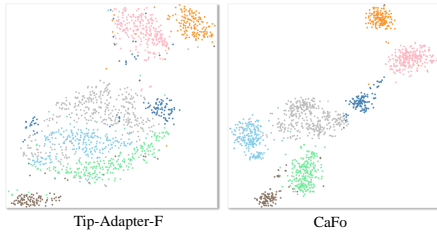
Figure 10. **t-SNE Visualization.** Different colors represent different categories on 16-shot ImageNet.
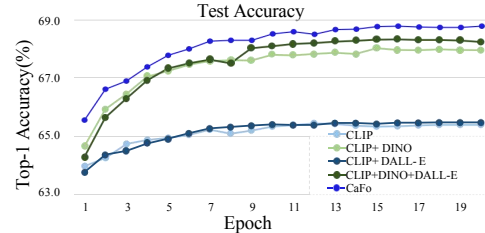
Figure 11. **Learning Curves of Test Accuracy (%)** for different combinations of pre-trained models on 16-shot ImageNet.

Our top prediction: great white shark

-With GPT-3 prompts:
-Score: 26.39

...is large, with a dark gray upper body and white underside.
...can be identified by its large size, wide-set eyes, and distinctive white belly.
...

-With CLIP templates:
-score: 26.02

a photo of the small [great white shark].
art of the [great white shark].
a bad photo of the [great white shark].
...

CLIP's top prediction: tiger shark

-With GPT-3 prompts:
-Score: 26.27

...are one of the largest shark species....
...are large, predatory sharks with a dark blue or grey back and white belly.
...

-With CLIP templates:
-Score: 26.52

a photo of the small [tiger shark].
art of the [tiger shark].
a origami [tiger shark].
a bad photo of the [tiger shark].
...

Our top prediction: tiger shark

-With GPT-3 prompts:
-Score: 26.33

...are one of the largest shark species....
...are large, predatory sharks with a dark blue or grey back and white belly.
...

-With CLIP templates:
-Score: 24.92

a photo of the small [tiger shark].
art of the [tiger shark].
a origami [tiger shark].
a bad photo of the [tiger shark].
...

CLIP's top prediction: great white shark

-With GPT-3 prompts:
-Score: 26.31

... are the largest species of shark in the world.
...looks like a large, bulky fish with a pointed nose, dark eyes, and a white underbelly.
...

-With CLIP templates:
-Score: 25.42

a photo of the small [great white shark].
art of the [great white shark].
a bad photo of the [great white shark].
...

Our top prediction: tiger shark

-With GPT-3 prompts:
-Score: 26.91

...are one of the largest shark species....
...are large, predatory sharks with a dark blue or grey back and white belly.
...

-With CLIP templates:
-Score: 26.08

a photo of the small [tiger shark].
art of the [tiger shark].
a bad photo of the [tiger shark].
...

CLIP's top prediction: hammerhead shark

-With GPT-3 prompts:
-Score: 26.63

...looks like a shark with a large head that resembles a hammer.
...looks like a shark with a wide, flat head that resembles a hammer.

-With CLIP templates:
-Score: 26.17

a photo of the small [hammerhead shark].
art of the [coucal].
a bad photo of the [hammerhead shark].
...

Our top prediction: stingray

-With GPT-3 prompts:
-Score: 29.20

...has a flat body and a long tail with a stinger on the end.
...is a large, flat fish with a long tail that has a sharp spine on the end of it.
...

-With CLIP templates:
-Score: 26.80

a photo of the small [stingray].
art of the [hen].
a origami [hen].
a bad photo of the [stingray].
...

CLIP's top prediction: electriv ray

-With GPT-3 prompts:
-Score: 29.03

...is a flat fish that can deliver a powerful electric shock.
...is a flat, disk-shaped fish that can grow up to two feet in length.
...

-With CLIP templates:
-Score: 27.05

a photo of the small [electric ray].
art of the [electric ray].
a bad photo of the [electric ray].
...

Figure 12. **Additional Visualization of GPT-3's Prompts for CLIP.** Above examples are from the ImageNet dataset.
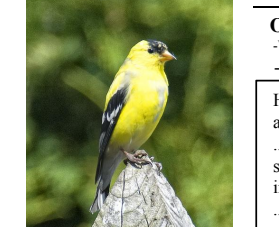
Our top prediction: hen

-With GPT-3 prompts:
-Score: 24.33

...are typically smaller and more delicate-looking than roosters. ...are small, domesticated birds that are typically considered female.
...

-With CLIP templates:
-Score: 21.61

a photo of the small [hen]. art of the [hen]. a origami [hen]. a [hen] in a video game. a bad photo of the [hen].
...

CLIP's top prediction: coucal

-With GPT-3 prompts:
-Score: 22.94

....is a crow-like bird with a long tail and a loud call. ...is a bird with a long tail and a dark brown plumage...
...

-With CLIP templates:
-Score: 23.30

a photo of the small [coucal]. art of the [coucal]. a origami [coucal]. a [coucal] in a video game. a bad photo of the [coucal].
...

Our top prediction: ostrich

-With GPT-3 prompts:
-Score: 29.14

...can be identified by their long necks, long legs, and wings. ...by their long necks and legs, their large egg-laying body, and their lack of wings.
...

-With CLIP templates:
-Score: 27.73

a photo of the small [ostrich]. art of the [ostrich]. a origami [ostrich]. a [ostrich] in a video game. a bad photo of the [ostrich].
...

CLIP's top prediction: bustard

-With GPT-3 prompts:
-Score: 28.97

...are a type of game bird with a heavy body and long legs. Large, long-necked bird with a big body and small head.
...

-With CLIP templates:
-Score: 27.89

a photo of the small [bustard]. art of the [bustard]. a origami [bustard]. a [bustard] in a video game. a bad photo of the [bustard].
...

Our top prediction: goldfish

-With GPT-3 prompts:
-Score: 24.92

Goldfish are small, orange fish with shiny scales. The easiest way to identify a goldfish is by its color.
...

-With CLIP templates:
-Score: 23.42

a photo of the small [goldfish]. art of the [goldfish]. a origami [goldfish]. a [goldfish] in a video game. a bad photo of the [goldfish].
...

CLIP's top prediction: coral reef

-With GPT-3 prompts:
-Score: 23.36

...is a type of biotic reef developing in tropical waters. ...a large underwater structure made up of many small stony coral polyps.
...

-With CLIP templates:
-Score: 23.53

a photo of the small [coral reef]. art of the [coral reef]. a bad photo of the [coral reef].
...

Our top prediction: house finch

-With GPT-3 prompts:
-Score: 26.84

House finches have red heads and red breasts. ...a small, plump songbird with a short tail and a wingspan of 8-9 inches.
...

-With CLIP templates:
-Score: 25.17

a photo of the small [ouse finch]. art of the [house finch]. a bad photo of the [house finch].
...

CLIP's top prediction: coucal

-With GPT-3 prompts:
-Score: 25.17

...a black bird with a long tail that is native to Africa ...a species of bird that is typically dark in color with a long tail.
...

-With CLIP templates:
-Overall score: 25.48

a photo of the small [coucal]. art of the [coucal]. a origami [coucal]. a [coucal] in a video game. a bad photo of the [coucal].
...

Figure 13. **Additional Visualization of GPT-3's Generated Prompts.** Above examples are from the ImageNet dataset.
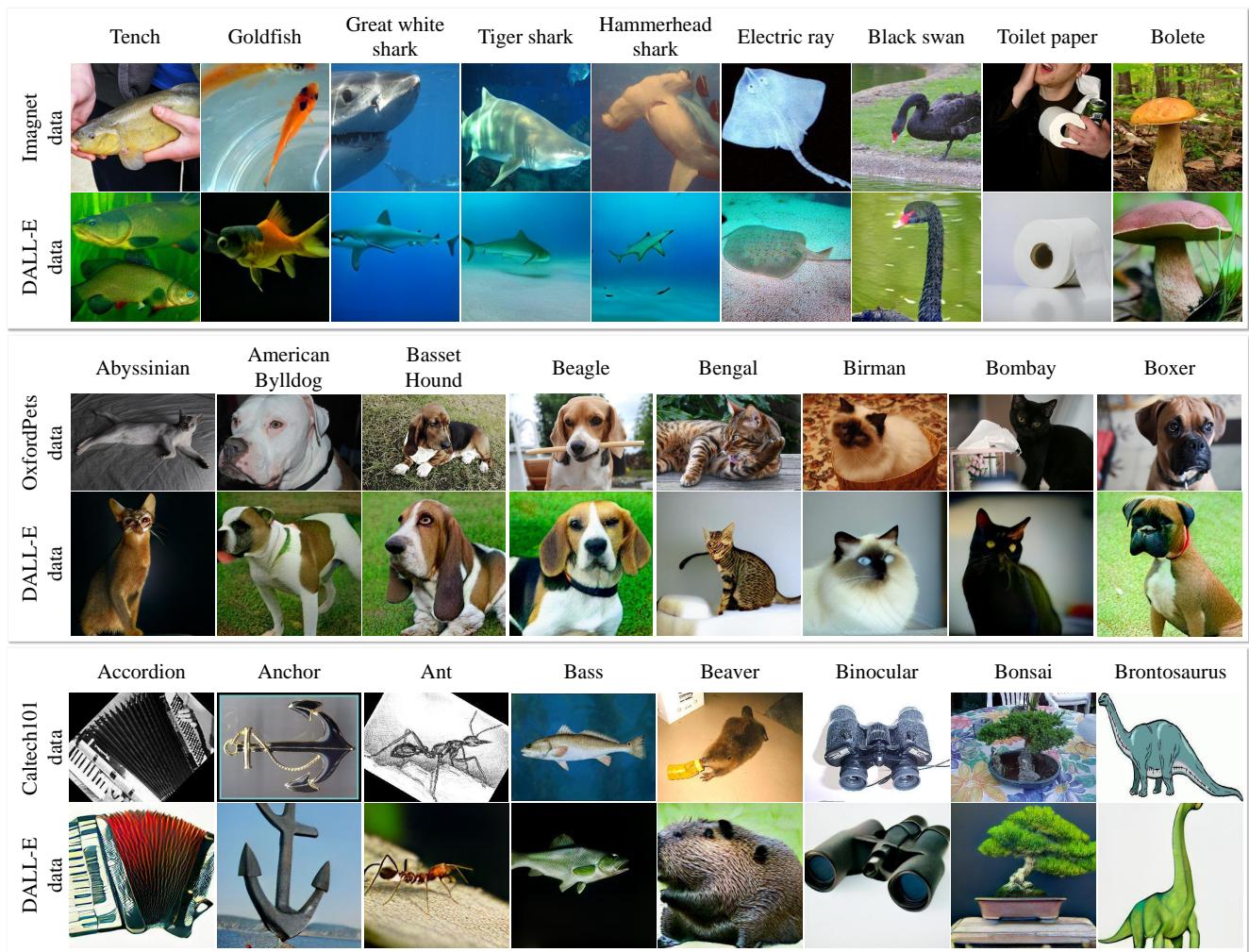
Figure 14. **Additional Visualization of DALL-E's Prompts for CLIP.**