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

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

[1]Shenzhen Institute of Advanced Technology, Chinese Academy of Science
[2]CUHK MMLab [3]Shanghai Artificial Intelligence Laboratory [4]CPII under InnoHK

{zhangrenrui, huxiangfei, gaopeng}@pjlab.org.cn,
hsli@ee.cuhk.edu.hk

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

1. Additional Performance Comparison

In Figure 1, we compare the performance of CaFo without DALL-E's [7] generated images or GPT-3's [1] created prompts on 10 datasets, which still consistently outperform the second-best Tip-Adapter-F.

2. Additional Ablation Study

Zero-shot DALL-E. We additionally show the ablation study of zero-shot generation by DALL-E on other three datasets in Table 2, 3 and 4. We explore the best synthetic number K' for each category of different shots. Same as the results on ImageNet, the larger K' does not lead to better few-shot performance since larger K' would contain more low-quality images and adversely affect the cache model.

CLIP's Visual Encoders. We conduct CaFo with different CLIP's [6] visual encoders for comparison with other methods. As shown in Table 1, CaFo consistently achieves leading performance with different visual backbones, indicating our generalizability to network architectures.

* Equal contribution. † Corresponding author

DALL-E	1	2	4	8	16
1	23.61	25.14	32.25	39.84	**49.05**
2	23.31	26.04	**32.94**	**40.38**	48.60
4	24.36	**26.13**	32.58	39.42	47.37
8	**24.96**	26.04	31.92	37.53	45.06
16	24.84	26.01	31.41	37.17	42.27

Table 2. **Zero-shot Results (%) on FGVCAircraft Dataset.**

DALL-E	1	2	4	8	16
1	67.51	**70.45**	72.54	77.80	79.51
2	67.91	69.1	72.54	77.16	**79.94**
4	68.09	70.21	**72.96**	**78.06**	79.75
8	**68.60**	69.36	71.37	76.74	79.43
16	67.78	68.91	71.90	76.47	78.88

Table 3. **Zero-shot Results (%) on UCF101 Dataset.**

DALL-E	1	2	4	8	16
1	**64.89**	**66.81**	**69.17**	**70.34**	**72.60**
2	64.70	66.63	69.08	70.33	72.26
4	64.70	66.46	68.62	70.09	72.25
8	64.16	65.62	68.23	69.46	71.78
16	64.03	65.75	67.19	69.29	70.97

Table 4. **Zero-shot Results (%) on SUN397 Dataset.**

Other Foundation Models. For the cache model, we investigate other pre-trained foundation models besides CLIP and DINO [2], including SimCLR [3], MAE [4], and SLIP [5]. We preserve the prompting and generation by

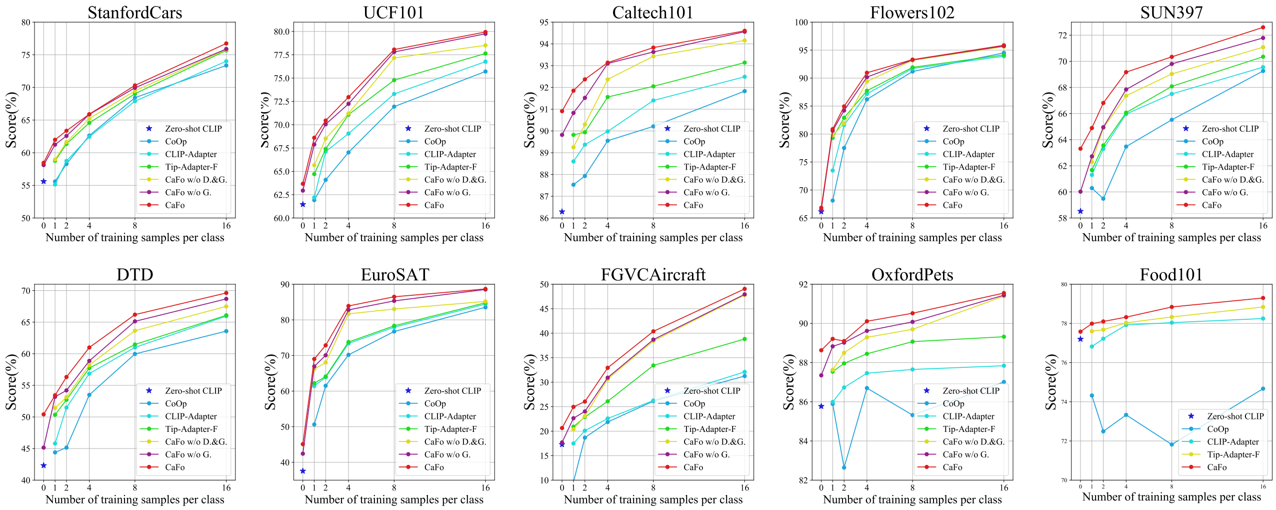

Figure 1. **Performance (%) Comparison on 10 Datasets.** Our method shows *state-of-the-art* performance for all few-shot settings on different datasets. 'CaFo w/o D.&G.' denotes CaFo without DALL-E's generated images and GPT3's created prompts.

Setting	ImageNet		OxfordPets		EuroSAT	
CLIP+SimCLR	62.3	65.7	87.1	89.4	55.7	75.9
CLIP+MAE	62.2	65.5	87.1	89.1	63.7	72.7
DINO+MAE	63.0	68.4	88.8	91.9	60.0	88.0
DINO+SimCLR	63.1	68.5	88.8	91.3	70.7	87.7
CLIP+DINO	63.8	68.8	89.2	91.6	69.0	88.7
SLIP+DINO	71.0	75.6	92.2	94.0	71.3	88.6

Table 5. **Ablation Study (%) of Other Foundation Models in the Cache Model.** We report the accuracy of 1 and 16 shots on ImageNet, OxfordPets, and EuroSAT.

GPT-3 and DALL-E, along with p_{ZS} as the ensemble baseline during adaptive inference. As shown in Table 5, 'CLIP+DINO', as our final solution, performs the best among different pre-training foundation models. Also, as an enhanced version of CLIP, SLIP can intuitively achieve higher accuracy in CaFo.

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 7, 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$\sim$16, and the larger number normally cannot get the better result, probably due to the low-quality synthetic images. On Caltech101 and EuroSAT, zero-shot CaFo largely surpasses CLIP by +4.62% and +7.54%, indicating our superiority under zero-shot settings.

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 6. **Ablation Study (%) of Hyperparameter** β**.** We report the 16-shot accuracy on ImageNet.

Hyperparameter β**.** In Formula 5 and 6, we utilize a non-linear 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 6, we experiment CaFo with different β on 16-shot ImageNet and observe 0.6 performs the best.

3. Additional Visualization

DALL-E's Generated Images. In Figure 4, 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.

GPT-3's Prompts for CLIP. In Figure 5 and 6, 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.

t-SNE. We present the t-SNE visualization of our CaFo and the second-best Tip-Adapter-F in Figure 2. CaFo shows more contrastive distribution of category clusters and well mitigates some aliasing between similar classes.

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 7. **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.

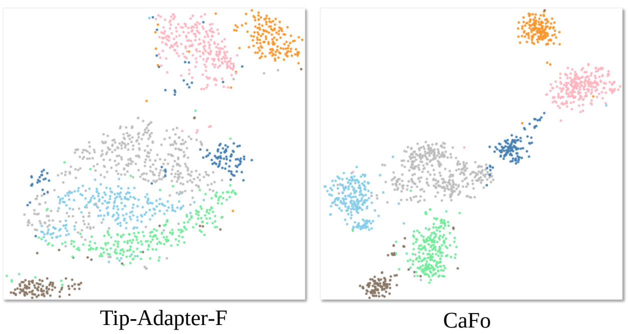

Figure 2. **t-SNE Visualization.** Different colors represent different categories on 16-shot ImageNet.

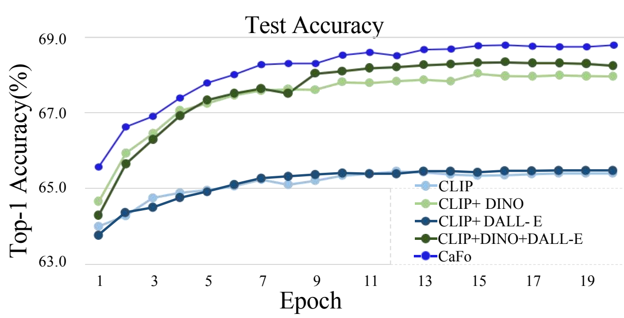

Figure 3. **Learning Curves** of Test Accuracy (%) for different combinations of pre-trained models on 16-shot ImageNet.

Learning Curves. In Figure 3, 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.

References

[1] 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. 1

[2] 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. 1

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1

[4] 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. 1

[5] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 529–544. Springer, 2022. 1

[6] 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

[7] 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. 1

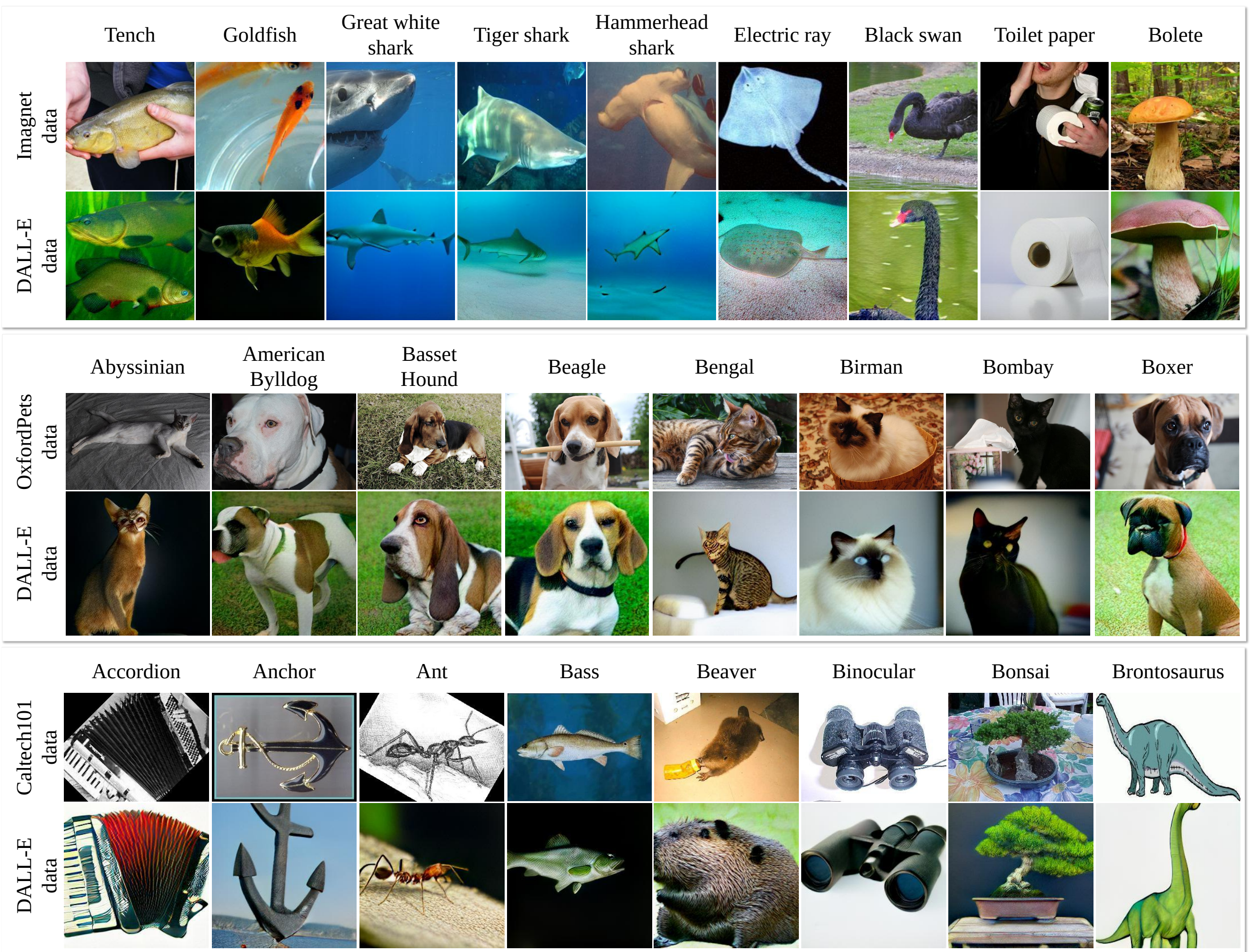

Figure 4. **Additional Visualization of DALL-E's Generated Images.** Examples are from ImageNet, OxfordPets and Caltech101 datasets.

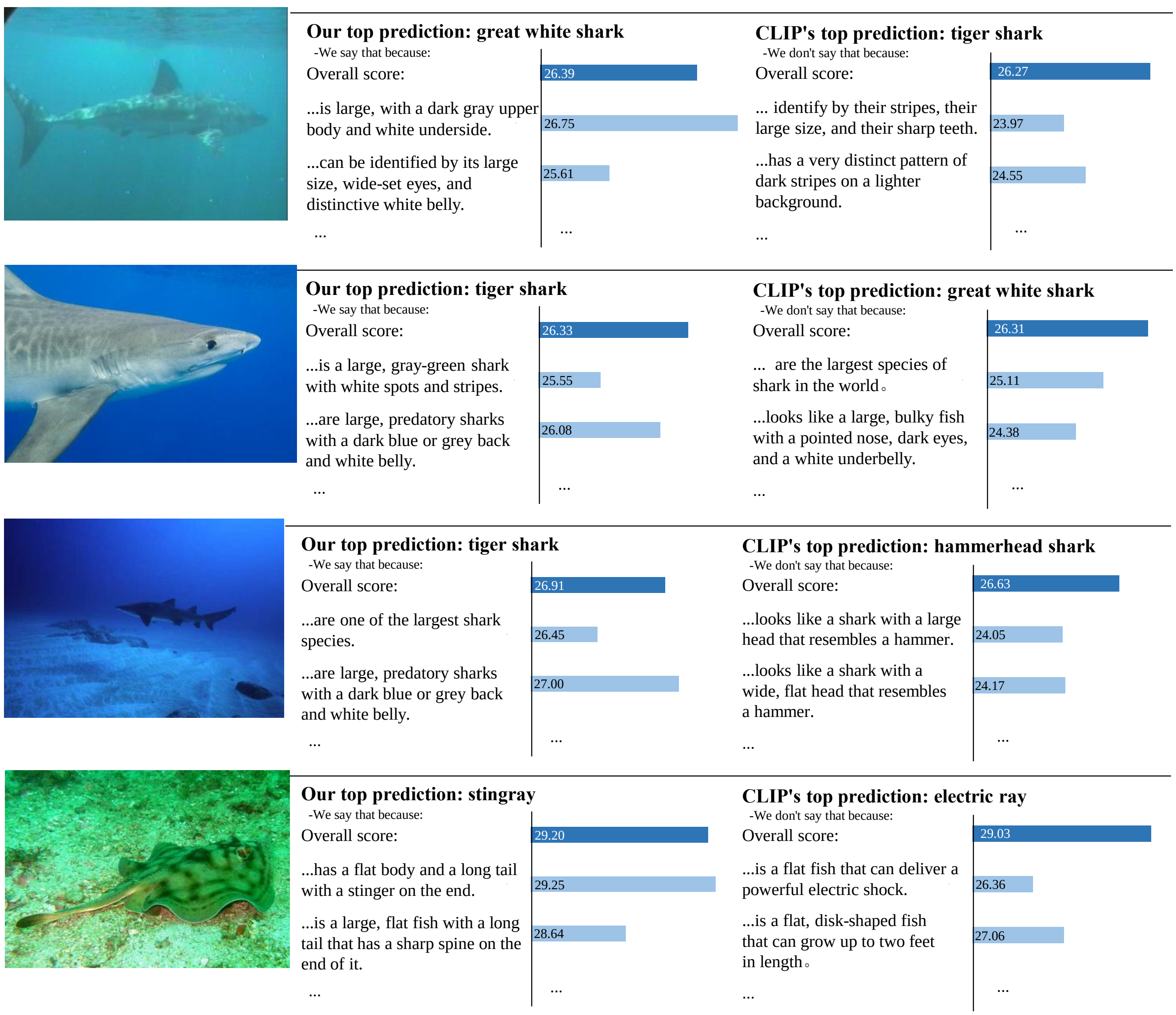

Figure 5. **Additional Visualization of GPT-3's Prompts for CLIP.** Above examples are from the ImageNet dataset.

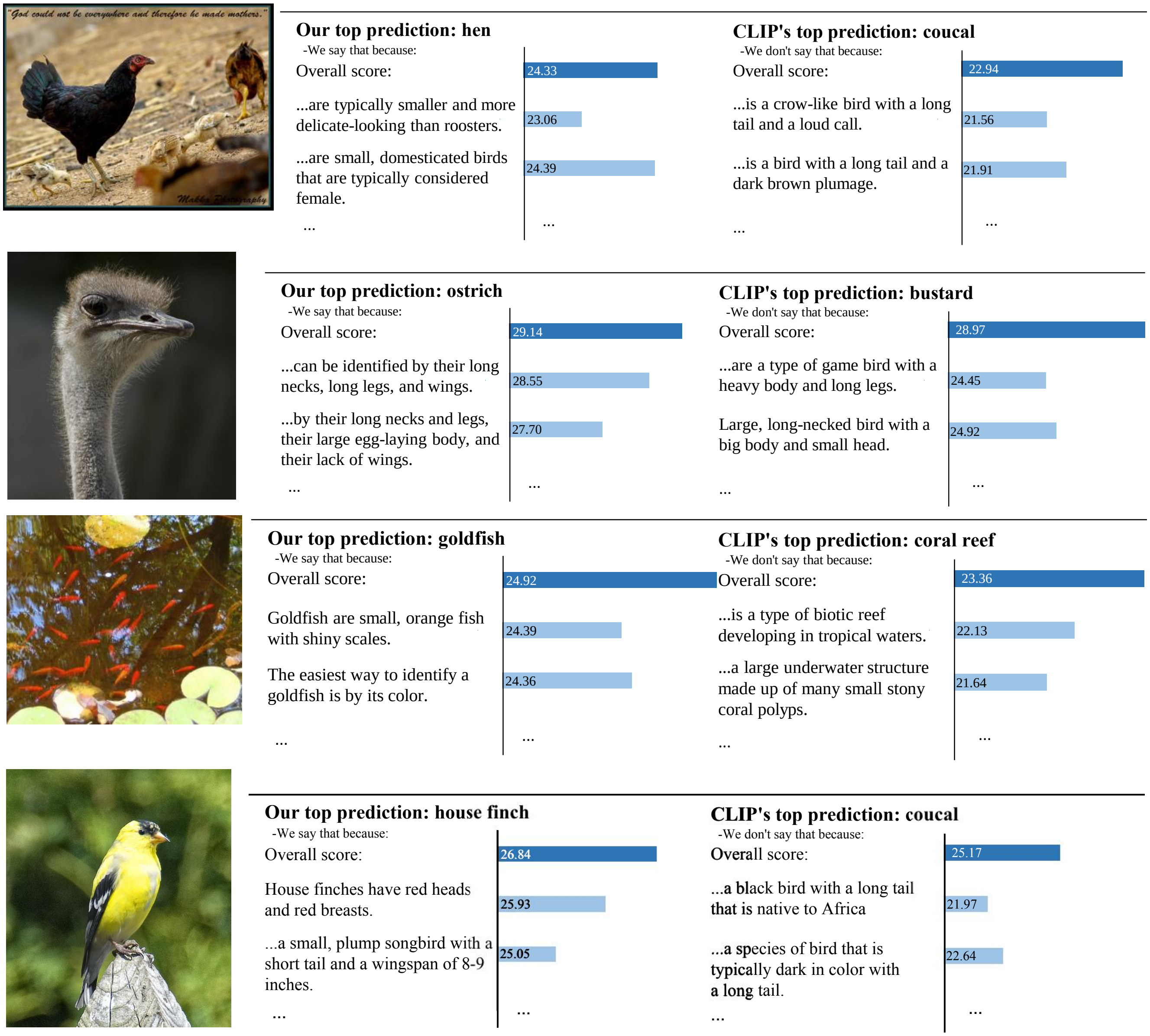

Figure 6. **Additional Visualization of GPT-3's Prompts for CLIP.** Above examples are from the ImageNet dataset.