SUPPLEMENTARY MATERIAL: Waffling around for Performance: Visual Classification with Random Words and Broad Concepts

In this supplementary material, we first provide a collection of additional results in §A which extend those presented in the main paper to more backbone models. Finally, we showcase the GPT-generated descriptors for our additionally used benchmarks beyond [36] (§B), and present some exemplary images from the eleven benchmarks used in this work in Fig. 7.

A. Additional results

Motivational experiments for random class descriptors. In Tab. 9, we extend our motivational experiments on random class descriptor assignment to motivate WaffleCLIP from Tab. 1, highlighting similar behaviour on both a larger ViT-L/14 and a ResNet50 backbone network. Descriptor randomization does not result in a significant drop in performance, but rather yields performances that match DCLIP.

Comparison of WaffleCLIP and DCLIP. Tab. 10 extends results from Tab. 2 on the ViT-L/14 and ResNet50 backbones, in which WaffleCLIP as a standalone method, as well as equipped with high-level concepts and/or joint usage of LLM-generated descriptors, is compared to DCLIP. The results confirm our conclusions drawn in §4.2, wherein WaffleCLIP, without access to any external LLM, can match the performance of LLM-descriptor-based approaches like DCLIP. In addition to that, we again find complementarity of randomized descriptors and LLM-generated descriptors. Furthermore, we observe performance gains through the usage of automatically generated high-level concepts.

Progression from systematic to fully randomized descriptor scrambling. Tab. 11 extends the descriptor scrambling progression studies from Tab. 4 to two additional backbones, namely, ViT-L/14 and ResNet50. Similar to the ViT-B/32 backbone, a move from systematic semantic shifts to independently subsampled descriptors can recover and even beat the performance of DCLIP.

B. Exemplary GPT-3 generated descriptors for additional benchmarks

As we introduce descriptions for three additional datasets beyond those used in [36], we provide four example descriptors for three random classes in each dataset.

Flowers102

Pink Primrose

- "delicate flower"
- "five petals in a star shape"
- "pink in color"
- "often has yellow center"

Balloon Flower

- "a delicate flower with five petals"
- "a unique balloon-like shape"
- "a star-shaped center in the middle of the flower"
- "vibrant colors such as pink, purple, blue, white, and yellow"

Sunflower

- "large, bright yellow petals"
- "a dark center surrounded by disk florets"
- "long stem"
- "a single, long, narrow leaves tapered to a point"

FGVCAircraft

A300

- "black or silver color"
- "a rectangular body with rounded edges"
- "two lens ports"
- "a mode dial"

EMB-120

- "a cabin with 30-33 seats"
- "a distinctive high-wing design"
- "two Pratt and Whitney PW118 turboprop engines"
- "a T-tail configuration"

Tornado

- "dark, rotating funnel-shaped cloud"
- "strong winds"
- "dark clouds"
- "heavy precipitation"

ViT-L/14	ImageNetV2	ImageNet	CUB200	EuroSAT	Places365	Food101	Oxford Pets	DTD	**Avg**
CLIP [47]	67.90	73.37	62.24	56.03	40.46	92.55	93.30	52.87	67.34
DCLIP [36]	**69.72**	75.26	63.53	58.72	**42.60**	92.81	93.89	**56.60**	69.14
DCLIP (same, 1x)	69.27 ±0.23	75.05 ±0.15	64.21 ±0.36	57.59 ±1.72	42.01 ±0.23	93.15 ±0.13	93.97 ±0.22	55.16 ±0.47	68.80 ±0.66
DCLIP (same, 2x)	69.58 ±0.21	**75.30** ±0.16	**64.30** ±0.26	**59.32** ±1.63	42.28 ±0.17	**93.31** ±0.05	**94.04** ±0.11	55.31 ±0.50	**69.18** ±0.62
ResNet50	ImageNetV2	ImageNet	CUB200	EuroSAT	Places365	Food101	Oxford Pets	DTD	**Avg**
CLIP [47]	51.34	58.16	45.20	28.09	36.63	78.37	83.76	38.51	52.51
DCLIP [36]	52.70	59.66	**47.76**	34.27	38.39	78.59	**85.77**	**41.01**	54.77
DCLIP (same, 1x)	52.63 ±0.28	59.69 ±0.30	**47.76** ±0.39	32.74 ±1.49	38.63 ±0.22	80.08 ±0.58	85.36 ±0.52	40.77 ±0.63	54.71 ±0.67
DCLIP (same, 1x)	**52.89** ±0.23	**59.90** ±0.26	47.70 ±0.29	**34.37** ±1.27	**38.93** ±0.21	**80.11** ±0.30	85.34 ±0.29	40.91 ±0.79	**55.02** ±0.58

Table 9: **Motivating random class descriptors - additional backbones.** Extension of our motivational experiments from Tab. 1 with ViT-L/14 and ResNet50 backbones.

ViT-L/14	ImageNetV2	ImageNet	CUB200	EuroSAT	Places365	Food101	Oxford Pets	DTD	**Avg**
CLIP [47]	67.90	73.37	62.24	56.03	40.46	92.55	93.30	52.87	67.34
+ Concepts	↓	↓	63.01	61.23	41.07	93.52	93.65	↓	68.32
DCLIP [36]	**69.72**	75.26	63.53	58.72	42.60	92.81	93.89	**56.60**	69.14
`WaffleCLIP` (ours)	69.48 ±0.08	75.30 ±0.04	**64.18** ±0.13	**61.17** ±0.35	42.26 ±0.10	93.31 ±0.09	91.98 ±0.11	53.94 ±0.29	68.95 ±0.18
+ Concepts	↓	↓	63.40 ±0.17	60.20 ±0.87	42.57 ±0.09	**93.65** ±0.05	**94.38** ±0.08	↓	69.12 ±0.33
+ GPT descr.	**69.80** ±0.13	**75.57** ±0.06	**64.32** ±0.21	**60.63** ±1.23	**42.96** ±0.12	93.28 ±0.08	93.35 ±0.22	**56.33** ±0.42	**69.53** ±0.48
+ GPT descr. + Concepts	↓	↓	63.14 ±0.16	**61.82** ±1.07	**42.95** ±0.09	93.49 ±0.04	94.12 ±0.09	↓	**69.65** ±0.42
ResNet50	ImageNetV2	ImageNet	CUB200	EuroSAT	Places365	Food101	Oxford Pets	DTD	**Avg**
CLIP [47]	51.34	58.16	45.20	28.09	36.63	78.37	83.76	38.51	52.51
+ Concepts	↓	↓	46.60	34.06	37.43	80.89	83.43	↓	53.80
DCLIP [36]	**52.70**	59.66	47.76	34.27	38.39	78.59	**85.77**	**41.01**	54.77
`WaffleCLIP` (ours)	**52.89** ±0.15	**60.12** ±0.12	47.68 ±0.15	31.34 ±0.47	38.32 ±0.10	79.68 ±0.17	84.32 ±0.20	39.25 ±0.27	54.20 ±0.23
+ Concepts	↓	↓	**48.34** ±0.13	35.08 ±0.42	39.03 ±0.08	**81.38** ±0.08	**85.80** ±0.12	↓	55.24 ±0.21
+ GPT descr. + Concepts	↓	↓	**48.41** ±0.21	**37.36** ±0.62	**39.43** ±0.07	81.17 ±0.09	**85.82** ±0.16	↓	**55.75** ±0.26

Table 10: **Performance of `WaffleCLIP` with additional backbones.** Here, we extend the comparison of `WaffleCLIP` (Tab. 2) to GPT-generated fine-grained class descriptors in DCLIP [36] for ViT-L/14 and ResNet50 backbones. We find similarly consistent insights, where our LLM-free `WaffleCLIP` can match the performance of DCLIP. Joint usage of both randomized and LLM-generated descriptors again reveals complementarity (`WaffleCLIP` + *GPT descr*). In addition to that, the usage of automatically extracted high-level semantic concepts can provide consistent additional performance gains (+ *Concepts*). We use (↓) to denote the same results as previous lines where high-level concept guidance is not applicable.

Stanford Cars

Acura TL Sedan 2012

- `"silver, grey, or black exterior"`
- `"Acura logo on the front grille"`
- `"distinctive headlights"`
- `"chrome accents on the exterior"`

BMW X6 SUV 2012

- `"four-door SUV"`
- `"sloping roof-line"`
- `"signature BMW kidney grille"`
- `"round headlights and taillights"`

Honda Odyssey Minivan 2012

- `"four doors and a hatchback"`
- `"a curved hood"`
- `"wide, round headlights"`
- `"a Honda logo"`

ViT-L/14	ImageNetV2	ImageNet	CUB200	EuroSAT	Places365	Food101	Oxford Pets	DTD	**Avg**
DCLIP [36]	**69.72**	**75.26**	63.53	**58.72**	**42.60**	92.81	93.89	**56.60**	**69.14**
DCLIP (interchanged)	66.44 ±0.12	72.07 ±0.15	63.62 ±0.44	51.49 ±4.89	37.06 ±0.41	91.30 ±0.30	93.74 ±0.28	49.84 ±0.78	65.69 ±1.77
DCLIP (scrambled)	68.68 ±0.21	74.47 ±0.11	63.78 ±0.13	55.98 ±2.01	41.29 ±0.23	92.29 ±0.20	93.52 ±0.18	53.28 ±1.12	67.91 ±0.83
DCLIP (random, 1x)	68.01 ±0.22	73.89 ±0.08	63.81 ±0.22	55.72 ±2.01	40.32 ±0.29	92.37 ±0.31	93.60 ±0.19	52.83 ±0.46	67.57 ±0.76
DCLIP (random, 5x)	69.27 ±0.17	75.11 ±0.08	64.25 ±0.16	58.34 ±1.55	42.11 ±0.14	93.22 ±0.12	93.88 ±0.09	55.28 ±0.23	68.93 ±0.57
ResNet50	ImageNetV2	ImageNet	CUB200	EuroSAT	Places365	Food101	Oxford Pets	DTD	**Avg**
DCLIP [36]	52.70	59.66	**47.76**	34.27	38.39	78.59	**85.77**	**41.01**	54.77
DCLIP (interchanged)	49.80 ±0.22	56.35 ±0.06	47.68 ±0.32	28.17 ±4.43	33.77 ±0.34	77.59 ±0.29	84.60 ±0.63	35.81 ±1.12	51.72 ±1.64
DCLIP (scrambled)	52.20 ±0.20	59.21 ±0.06	47.60 ±0.39	34.98 ±2.00	37.90 ±0.18	78.33 ±0.14	85.07 ±0.34	39.19 ±0.95	54.31 ±0.81
DCLIP (random, 1x)	51.60 ±0.29	58.29 ±0.15	47.37 ±0.23	30.18 ±4.18	36.82 ±0.26	78.87 ±0.24	84.52 ±0.17	38.89 ±0.85	53.32 ±1.52
DCLIP (random, 5x)	52.81 ±0.09	59.73 ±0.05	47.74 ±0.10	34.53 ±0.74	38.62 ±0.15	80.20 ±0.13	85.30 ±0.15	40.29 ±0.46	54.90 ±0.32

Table 11: **Progression from systematic to fully randomized descriptor scrambling - additional backbones.** We extend our descriptor scrambling progression studies from Tab. 4 to two additional backbones: ViT-L/14 and ResNet50. In both cases, the same trend can be seen, in which a move from systematic semantic shift to independently subsampled descriptors can recover the performance of DCLIP after an initial performance drop.

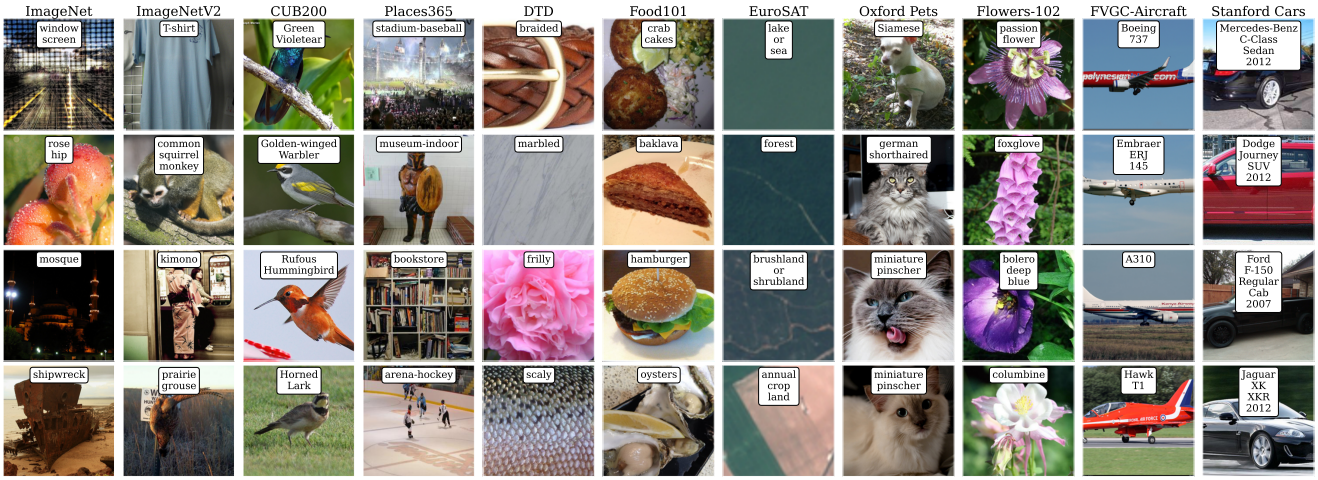

Figure 7: To get an intuition of the different visual classification tasks, we showcase samples of four randomly selected classes for each of the eleven utilized visual classification benchmarks.