Supplementary Material for "Enhancing Visual Classification using Comparative Descriptors"

A. Hyperparameters

As mentioned in our main paper, we utilize two hyperparameters: n represents the number of similar classes for each class, and k represents the maximum number of descriptors remaining after filtering. We reveal the hyperparameters used in our experiment in Table A. Note that since the SAT dataset has only 10 classes, each class can have up to 9 similar classes.

	IN1k	**IN1k**-V2	**Caltech**	**CIFAR**	**CUB**	**SAT**	**Places**	**Food**	**Pets**	**DTD**	**Flowers**	**Aircraft**	**Cars**
n	10	10	10	10	10	9	10	10	10	10	10	10	10
k	10	10	10	20	5	10	15	20	20	5	5	15	15

Table A. Hyperparameters used for each dataset in our experiments.

B. Image Classification With Additional Backbones

We conduct additional experiments using ViT-L/14 and ResNet50 to further demonstrate the robustness of our method across different model architectures and configurations, complementing the ViT-B/32 used in the main paper. These results indicate the consistent performance and adaptability of our approach regardless of the underlying model. The results are presented in Table B and Table C, showing that our approach consistently outperforms existing methods.

ViT-L/14	**IN1k**	**IN1k-V2**	**Caltech**	**CIFAR**	**CUB**	**SAT**	**Places**	**Food**	**Pets**	**DTD**	**Flowers**	**Aircraft**	**Cars**	**Avg**
CLIP	73.36	67.92	87.05	77.17	62.32	56.05	40.45	92.55	93.30	52.87	74.73	44.91	76.50	69.17
DCLIP	75.92	70.00	89.89	78.84	65.33	63.19	42.84	93.39	93.57	54.95	78.68	49.80	76.32	71.75
WaffleCLIP	75.36	69.55	90.00	79.04	64.19	60.60	42.19	93.09	91.79	53.72	76.35	48.91	76.68	70.88
Ours	76.04	70.48	90.69	79.71	65.77	62.07	43.83	93.54	94.55	57.87	79.15	51.07	76.69	72.42
Ours + Filtering	76.77	70.59	91.45	79.80	63.76	62.80	44.00	93.42	93.79	62.83	80.54	50.34	75.82	72.76

Table B. Image classification accuracy (%) with ViT-L/14.

RN50	**IN1k**	**IN1k-V2**	**Caltech**	**CIFAR**	**CUB**	**SAT**	**Places**	**Food**	**Pets**	**DTD**	**Flowers**	**Aircraft**	**Cars**	**Avg**
CLIP	58.15	51.37	75.96	42.07	45.29	28.12	36.70	78.38	83.70	38.62	61.34	21.21	54.00	51.92
DCLIP	60.56	53.28	79.99	43.32	49.09	30.19	39.75	79.15	84.33	41.97	66.95	22.86	53.76	54.25
WaffleCLIP	60.22	53.04	78.94	43.50	47.60	30.69	38.24	79.36	84.08	39.34	65.11	23.60	53.61	53.64
Ours	60.83	53.26	80.14	44.02	48.93	34.19	40.38	79.85	86.29	43.40	67.25	24.36	54.32	55.17
Ours + Filtering	61.71	53.79	81.78	44.00	49.24	34.51	40.56	80.67	84.90	49.53	69.33	24.61	53.43	56.00

Table C. Image classification accuracy (%) with ResNet50.

C. Evaluation of Descriptor Quality through Question-Answering Task

To assess the quality of descriptors generated by our method, we conduct a question-answering (QA) task using LLaMA-3.1-Instruct, an open-source language model. For each class, we input its descriptor alongside three similar classes and ask the model to identify the best match. All hyperparameters remain at their default values. To ensure accurate results, we repeat the experiment five times and present the average outcome in Table D. Note that WaffleCLIP uses random strings as descriptors, making it unsuitable for this task. Descriptors generated by our method significantly outperform those from DCLIP, highlighting their superior quality.

LLaMA-3.1	IN1k	IN1k-V2	Caltech	CIFAR	CUB	SAT	Places	Food	Pets	DTD	Flowers	Aircraft	Cars	Avg
DCLIP	81.64	81.76	91.52	92.80	54.50	90.00	**90.47**	91.09	63.24	69.79	74.90	64.86	79.69	78.94
Ours	**92.34**	**90.76**	**98.21**	**96.00**	**85.60**	**100.00**	90.14	**96.63**	**91.89**	**94.47**	**89.61**	**99.43**	**96.63**	**93.98**

Table D. Question-answering accuracy (%) evaluated using LLaMA-3.1-8B-instruct.

D. Experimental results with high-level concepts

We show the classification results with high-level concepts, proposed in WaffleCLIP, in Table E. Note that we exclude datasets with more than one high-level concept, as in WaffleCLIP; for example, ImageNet (IN1k) is excluded because it contains many different kinds of objects, animals, food, etc. As shown in the results, our method consistently outperforms the baselines.

ViT-B/32	CUB	SAT	Places	Food	Pets	DTD	Flowers	Aircraft	Cars	Avg
CLIP + Concepts	52.12	40.09	39.30	84.67	86.81	42.71	64.37	26.94	59.37	55.15
DCLIP + Concepts	53.61	48.13	41.71	84.56	89.26	43.35	69.10	28.08	59.45	57.47
WaffleCLIP + Concepts	52.83	48.58	41.02	84.91	88.29	42.32	67.21	28.35	59.67	57.02
WaffleCLIP + Concepts + DCLIP descr.	53.25	48.67	41.82	85.00	88.58	43.93	68.19	28.56	59.77	57.53
Ours + Concepts	53.81	51.40	42.24	84.92	88.88	44.63	68.50	29.43	59.77	58.18
Ours + Filtering + Concepts	56.22	58.40	44.38	85.14	87.11	51.11	73.89	29.02	59.39	60.52

Table E. Image classification accuracy (%) with high-level concepts.

E. Inference Time

Our method introduces additional steps beyond the baselines, raising concerns about inference time. To address this, we measure inference time, which remains nearly identical across methods since actual classification uses pre-generated descriptors. We define inference time as the duration of steps before classification. Since descriptor generation with the large language model is shared by both methods and varies in API response time, we exclude it from the calculation, along with the one-time model load time ($\sim$2.3 seconds). We measure the additional inference time required by our steps—identifying similar classes and filtering—compared to the baseline, as shown in Table F. The increase in time is minimal relative to performance gains. This experiment, conducted on a single RTX 3090 GPU, has room for runtime optimization, suggesting potential further reductions in time.

ViT-B/32	IN1k	IN1k-V2	Caltech	CIFAR	CUB	SAT	Places	Food	Pets	DTD	Flowers	Aircraft	Cars
Similar	0.29	0.29	0.07	0.03	0.05	0.01	0.10	0.03	0.01	0.01	0.03	0.02	0.06
Filtering	28.93	29.07	7.27	3.23	5.89	0.71	14.14	3.12	1.49	1.41	3.04	2.43	7.03

Table F. The measured inference time (**s**). "Identifying similar classes" is denoted as "Similar".

F. Qualitative Results

We present the qualitative results in Table G to show the quality of the descriptors generated by our method and DCLIP. Our method minimizes features that are difficult for VLMs to leverage (e.g. numerical information) and provides more detail.

Dataset	Class	Ours	DCLIP
IN1k & IN1k-V2	drumstick	narrow, cylindrical shape rounded tip for striking surfaces long, slender wooden or plastic shaft tapered end or tip designed for striking uniform thickness along most of its length	smooth surface cylindrical shape often found in pairs relatively lightweight length approximately 15-18 inches (varies)
Caltech	golf-ball	dimpled texture Small, spherical shape Perfectly round shape without stitching solid white or occasionally other solid colors Presence on a grassy or golf-course background	small, spherical object typically white or sometimes colored covered in dimples (small indentations) may have brand names or logos printed on it standard size approximately 1.68 inches (42.67 mm) in diameter
CIFAR	house	Garage or driveway Surrounding lawn or garden Driveway or pathway leading to an entrance Residential surroundings such as nearby houses or streets Residential design and materials (e.g., brick, wood, or siding)	garage (optional) a chimney (optional) a building structure windows with glass panes eaves or gutters along the roofline
CUB	Downy Woodpecker	Vertical clinging posture on trees Long, chisel-like bill for pecking wood Black and white checkered pattern on wings Relatively small, pointed beak designed for pecking wood Short and stiff tail feathers used for support against tree trunks	White underparts Black and white plumage Black wings with white spots Black and white striped head Small size, about 6-7 inches in length
SAT	residential buildings or homes or apartments	Balconies or patios with furniture or plants Fences or boundaries surrounding the property Personal vehicles like cars or bicycles parked outside Urban or suburban layout with interconnected structures Presence of driveways and parking areas adjacent to buildings	doors a roof windows chimneys balconies or porches
Places	picnic_area	picnic tables grills or barbecues Wooden or metal picnic tables shade structures (e.g., pavilions or gazebos) recreational equipment like frisbees or balls	Trash bins Open grassy areas Trees or natural shade Signs indicating picnic areas People sitting or eating outdoors
Food	peking duck	sliced duck meat whole roasted duck crispy, caramelized skin thin pancakes or steamed buns for wrapping garnished with sliced cucumbers and scallions	Orange bill White plumage Rounded body shape Orange legs and webbed feet Garnished with sliced cucumbers and green onions
Pets	havanese	Dropped ears small, sturdy build Longer, wavy or curly coat longer ears that hang down, covered in long fur Tightly curled tail that often arches over the back	floppy ears small breed dog long, wavy or curly coat friendly and alert expression sturdy body with a height ranging from 8.5 to 11.5 inches
DTD	striped	parallel lines no intersecting lines lines of consistent color Consistent direction of lines uniform spacing between lines	It seems there might be a typo in your request. \"Striped\" is an adjective and needs a noun to be complete (e.g., striped shirt, striped animal). Could you please clarify what you are referring to with \"striped\"?"
Flowers	fritillary	downward-facing blooms narrow, lance-shaped leaves checkered or spotted floral patterns Distinctive pattern on the wings (if a butterfly) Bell-shaped, downward-facing flowers (if a flower)	butterfly or moth lance-shaped or linear leaves checkered or spotted underwings proboscis for feeding on nectar or other fluids wings often patterned with black spots or stripes
Aircraft	Cessna 172	Fixed landing gear Four-seat configuration Smaller tail and horizontal stabilizers single engine with a propeller in the front Enclosed cockpit with side-by-side seating	fixed landing gear four-seat configuration windows along the fuselage small, single-engine aircraft tricycle landing gear with a nose wheel
Cars	Chevrolet TrailBlazer SS 2009	SUV body style Lower, sportier stance SS badging on the grille and rear Integrated front bumper with fog lights Signature bowtie emblem on the front grille	SUV body style 20-inch aluminum wheels Wide, aggressive stance 6.0-liter V8 engine visible if the hood is open Interior features such as sport seats and specific SS embroidery

Table G. Qualitative results for all evaluated datasets. Since IN1k and IN1k-V2 have the same class, we mark them as the same.