HQ-CLIP: Leveraging Large Vision-Language Models to Create High-Quality Image-Text Datasets and CLIP Models

Supplementary Material

1. Experiments

DataComp Scale	Small	Medium	Large
CommonPool size	12.8M	128M	1.28B
Original DFN size	-	19.2M	192M
Reproduced DFN size	1.47M	14.7M	147M
Model	ViT-B/32	ViT-B/32	ViT-B/16
Batch size	4096	4096	8192

Table 1. Training setup and dataset scale.

1.1. Setup

Our experimental setup primarily follows the configuration established in DFN [1]. The original DFN methodology processes CommonPool datasets (12.8M/128M/1.28B) to derive filtered subsets of 1.92M (small), 19.2M (medium), and 192M (large) image-text pairs. Due to partial URL inaccessibility, we obtained reduced subsets of **1.47M (small)**, **14.7M (medium)**, and **147M (large)** pairs for our implementation. In model training, we strictly adhere to DFN's architectural specifications and batch size configurations. Notably, for the XLarge-scale model training, we employed CLIPA [4] to optimize computational efficiency and accelerate training convergence.

1.2. Ablation study

Ablation study on hard-negative sample quantity. We investigate the optimal number of hard-negative variants per image for identification tasks. As Table 8 demonstrates, empirical evidence suggests the single-sample configuration emerges as optimal. Although increasing the number of samples initially appears to benefit performance metrics, practical constraints such as prohibitive GPU memory demands and computational overhead prevent further scaling. Consequently, we select one hard-negative instance as the computationally efficient yet effective solution.

Ablation on the number of classes. Our framework employs a frequency-based selection of the top-K most prevalent tags from the VLM-generated tag repository. As empirically validated in Table 3, we systematically determine the optimal class quantity parameter K.

Ablation study on loss hyperparameters α and β. Performance sensitivity to the hard-negative identification loss weight (α) and short tag classification loss weight (β) is quantified in Tables 5 and 4. The optimal configuration is observed at $\alpha = 0.5$ and $\beta = 10$, where both loss components contribute maximally to model effectiveness.

Scale	Methods	Attribution	Relation
Medium	DFN†	54.2	53.2
	Ours	**61.1**	**54.4**
Large	DFN†	55.1	47.2
	Ours	**65.1**	**61.3**

Table 2. Comparison of attribution and relation metrics in the ARO benchmark [5].

Number of classes	3000	10000	30000	90000
ImageNet	39.2	**40.8**	40.5	40.6
ImageNet-Shifts	31.1	**32.9**	32.8	32.8
VTAB	40.3	40.5	40.5	**42.3**
Retrieval	36.6	37.4	**37.7**	37.3
Average over 38 datasets	39.9	40.2	40.1	**40.5**

Table 3. Ablation study on the number of classes.

β	1	10	100	1000
ImageNet	40.1	40.0	**40.7**	40.6
ImageNet-Shifts	32.4	32.4	**33.2**	32.1
VTAB	44.1	44.3	**44.7**	44.1
Retrieval	37.2	36.9	**37.5**	36.8
Average over 38 datasets	40.1	39.9	**40.5**	40.2

Table 4. Ablation study on the weight of $\mathcal{L}_{STC}$.

α	0.1	0.2	0.5	1
ImageNet	40.6	**40.7**	40.1	40.2
ImageNet-Shifts	**32.7**	32.5	32.5	32.2
VTAB	40.8	41.2	**41.6**	41.3
Retrieval	38.7	37.7	**38.1**	**38.1**
Average over 38 datasets	40.1	39.9	**40.7**	40.0

Table 5. Ablation study on the weight of $\mathcal{L}_{HNI}$.

1.3. Comparison with state-of-the-art method

ARO benchmark evaluation. As shown on Tab. 2, our approach exhibits superior comprehension of attribution and relation compared to the DFN† baseline. By benefiting from descriptions with enhanced semantic richness and the specialized hard-negative identification loss during training, our method achieves significant and scalable performance improvements on Visual Genome attribution metrics.

Comparison with VeCLIP. Given the exceptional performance claims of VeCLIP[3] in its original publication, com-

	Dataset size	IN	INv2	COCO	Flickr	Caltech101	CIFAR100	SVHN	DTD	OxPet	Flowers102	EuroSAT	RESISC45	Camelyon	Average
VeCLIP [3]	200M	64.6	57.7	**57.8**	**83.7**	83.1	68.1	44.9	**62.0**	72.6	68.5	47.4	55.1	**62.6**	62.7
Ours	148M	**70.6**	**63.1**	52.2	77.9	**93.1**	**81.0**	**45.9**	51.5	**89.5**	**69.0**	**47.6**	**60.6**	46.2	**64.9**

Table 6. Comparison of Our Method with VeCLIP. The metrics for VeCLIP are sourced from the original paper. Our method demonstrates superior average performance.

	IN	IN-Shifts	VTAB	Retrieval	Average over 38 datasets
VeCLIP*	52.5	45.9	46.8	55.2	48.2
Ours	**70.6**	**57.2**	**57.6**	**60.9**	**58.6**

Table 7. Comparison of Performance on the DataComp [2] Benchmark with VeCLIP. The metrics for VeCLIP were obtained by using the weights provided in its official GitHub repository, trained on the 100 Million dataset, and evaluated using the DataComp benchmark code and Hugging Face tools.

N^-	1	2	3	4
ImageNet	39.9	39.8	**40.0**	39.5
ImageNet-Shifts	**32.5**	**32.5**	32.1	32.3
VTAB	41.8	**42.0**	40.9	41.1
Retrieval	**38.1**	38.0	37.7	37.9
Average over 38 datasets	40.1	**40.2**	40.1	39.9

Table 8. Ablation study on number of hard negative samples M.

prehensive benchmarking becomes imperative. However, since VeCLIP did not include DataComp benchmark results in their work, a direct comparison in our main results table (Table 2) proves infeasible. We therefore provide supplementary comparisons with more performance metrics in the Supplementary Materials between our method and the ViT-B variant of VeCLIP trained on 200 million samples (as reported in their paper), where our approach demonstrates superior comprehensive performance (Table6).

To facilitate rigorous benchmarking, we sought to evaluate VeCLIP under the DataComp[2] framework. While the authors provide clear instructions for loading their ViT-H weights, documentation gaps were identified regarding ViT-B weight implementation. Technical challenges emerged from (1) framework-specific implementation details in TensorFlow and (2) compatibility constraints with VeCLIP's text encoder architecture in the DataComp library. To address these methodological challenges, we re-implemented a PyTorch version of VeCLIP's data pipeline and modified the DataComp evaluation code.

Due to technical limitations in loading VeCLIP model weights trained on the 200M subset, our analysis employs the 100M variant for standardized DataComp benchmark comparisons (Table 7). HQ-CLIP significantly outperforms VeCLIP. We are actively seeking verification through direct communication with the authors' team to ensure correct comparison and sincerely welcome their insights.

Figure 1. Comparison of recognition results between our model and DFN.

1.4. Recognition Results

Figure 1 shows the classification results of our model compared to the DFN model. For each image, binary classification is performed using manually crafted text to demonstrate the fine-grained understanding capability of the models. Our model shows better recognition of detailed semantics in the images.

1.5. Details of other experiments

We showcase the full 38 dataset result for some experiments on main paper, as shown in Tab. **??** and 10.

2. VLM-150M

2.1. Examples

We present some examples from the acquired dataset. As shown in Figure 2, we obtained more comprehensive annotations.

(a)

Description $\{d^+\}$:
The image shows a **backyard** paver patio area connected to a **concrete** slab, there is an **outdoor fire pit** made of brick in the center of the patio, the patio is surrounded by landscaping with trees and a wooden garden fence, the scene appears to be taken during the **day** with **wet** pavement.

Negative Description $\{d^-\}$:
The image shows an **indoor living room** with a **wooden** deck, there is **an indoor fireplace** made of brick in the center of the room, the floor is carpeted with a patterned design, the scene appears to be taken during the **night** with **dry** wooden flooring.

Negative Tags $\{t_i^-\}$:
indoor living room, wooden deck, garage, indoor fireplace, carpeted floor, kitchen area

Tags $\{t_i^+\}$:
backyard, paver patio, concrete slab, outdoor fire pit, landscaping, patio area, garden fence, trees

(b)

Description $\{d^+\}$:
The image shows a **double bedroom** with a river view, the bed has **white** linens with a **black and white blanket** at the foot, the room features a **wooden** bed frame with intricate carvings, there are nightstands on either side of the bed, each with a lamp, **a window with wooden** shutters is partially open, offering a view of the outdoors.

Negative Description $\{d^-\}$:
The image shows a **single bedroom** with a city view, the bed has **colorful** linens with a **blue and green blanket** at the foot, the room features a **metal** bed frame with simple design, there are no nightstands or lamps in the room, there are **no windows**, giving a closed-off feel to the room.

Negative Tags $\{t_i^-\}$:
single bedroom, city view, bed with colorful linens, blue and green blanket, metal bed frame, no nightstands, no lamps, no windows

Tags $\{t_i^+\}$:
double bedroom, river view, bed with white linens, black and white blanket, wooden bed frame, nightstands, lamps, window with curtains

(c)

Description $\{d^+\}$:
The image shows the **waterfront** of Long Beach, California, **a marina** with numerous boats is visible in the foreground, the background features modern buildings and **palm trees**, an **American flag** is prominently displayed in the center of the image, the scene is captured on a **clear day** with blue skies.

Negative Description $\{d^-\}$:
The image shows the **skyline** of New York City, a view of **the Empire State Building** is visible in the foreground, the background features urban skyscrapers and **a mountain range**, a **European flag** is prominently displayed in the center of the image, the scene is captured on a **cloudy day** with overcast skies.

Negative Tags $\{t_i^-\}$:
New York, New York City, Empire State Building, urban skyscrapers, mountains, forest, European flag

Tags $\{t_i^+\}$:
Long Beach, California, architecture, marina, boats, cityscape, waterfront, American flag

(d)

Description $\{d^+\}$:
The image depicts **a statue of Jesus** crucified on a cross inside a church, the scene is set during **the Easter Triduum**, with a focus on the religious significance of the cross, the background includes **a wooden altar** and other religious decorations, candles are visible, adding to the serene atmosphere of the church interior.

Negative Description $\{d^-\}$:
The image depicts **a statue of Buddha** in an outdoor setting, the scene is set during a **Christmas celebration**, with a focus on the festive decorations, the background includes **a metal altar** and secular decorations, candles of peace are visible, adding to the festive atmosphere of the outdoor setting.

Negative Tags $\{t_i^-\}$:
Buddha statue, outdoor setting, Christmas celebration, secular symbol, statue of liberty, candles of peace, metal altar

Tags $\{t_i^+\}$:
Jesus crucified, church interior, Easter Triduum, religious symbol, cross, statue of Jesus, candles, wooden altar

(e)

Description $\{d^+\}$:
The image shows a winding road leading up a **mountain** with a **cloudy sky**, there are **several cars** and trees visible along the road, the sky has **a mix of blue and green hues with some clouds**, an electric wire is prominently visible in the foreground, the scene captures **a natural landscape** with a sense of vast openness.

Negative Description $\{d^-\}$:
The image shows a **beach** with a **clear sky** and sunny weather, there are **bicycles** and people walking along the beach, the sky is clear with **no clouds**, an underpass is prominently visible in the foreground, the scene captures **an urban area** with a sense of bustling activity.

Negative Tags $\{t_i^-\}$:
beach, forest, sunny day, bicycles, urban area, clear sky, underpass, cityscape

Tags $\{t_i^+\}$:
mountain, road, cloudy sky, cars, landscape, nature, electric wire, scenic view

(f)

Description $\{d^+\}$:
The image shows a **small, white fluffy Shih Tzu dog** standing on a **wooden** surface, the dog has a **happy** expression with its tongue out, the background is an indoor setting with blurred furniture and a person, the Shih Tzu has **a round face and fluffy fur**.

Negative Description $\{d^-\}$:
The image shows a **large, black Golden Retriever** standing on a **grassy** surface, the dog has a **sad** expression with its head down, the background is an outdoor setting with trees and grass, the Golden Retriever has **a long face and short fur**.

Negative Tags $\{t_i^-\}$:
Golden Retriever, cat, wild animal, black dog, large breed, outdoor setting, sad dog, wild animal

Tags $\{t_i^+\}$:
Shih Tzu, dog, pet, white fluffy dog, small breed, indoor setting, happy dog, toy dog

(g)

Description $\{d^+\}$:
The image shows **the Domaine Saint-Clair Le Donjon in Etretat, France**, the scene features impressive **white cliffs** and rock formations extending into the sea, the water is a clear blue, creating a striking contrast with the chalk cliffs, the sky is **clear** with a hint of clouds, suggesting a **sunny day**.

Negative Description $\{d^-\}$:
The image shows **the Grand Canyon in Arizona, USA**, the scene features vast **desert** landscapes with sand dunes, the water is a river, creating a winding path through the mountains, the sky is **overcast** with no visible clouds, suggesting a **cloudy day**.

Negative Tags $\{t_i^-\}$:
Grand Canyon, Arizona, desert, mountains, urban landscape, river, USA

Tags $\{t_i^+\}$:
Domaine Saint-Clair Le Donjon, Etretat, cliffs, rock formations, coastal landscape, sea, France

(h)

Description $\{d^+\}$:
The image depicts a painting of **water lilies** on a pond by **Claude Monet**, the artwork is from the period **1897-1899**, the painting shows lily pads and two fully bloomed water lilies, the background is composed of **soft, swirling brushstrokes** with shades of blue and green.

Negative Description $\{d^-\}$:
The image depicts a painting of a **night sky** with swirling patterns by **Vincent van Gogh**, the artwork is from **the 19th century**, the painting shows a night sky filled with stars and a crescent moon, the background is composed of **abstract shapes** and vibrant colors.

Negative Tags $\{t_i^-\}$:
Vincent van Gogh, Starry Night, 19th century, abstract painting, night sky, stars, swirling patterns, landscape

Tags $\{t_i^+\}$:
Claude Monet, Water Lilies, 1897-1899, impressionist painting, water garden, lily pads, flowers, pond

Figure 2. Examples of VLM-150M.

Model	XCom2	LLaVA	Qwen2-VL	Qwen2-VL	Qwen2-VL	Qwen2-VL
Parameters	7B	7B	7B	2B	72B	7B
GPT4o SFT	✓	✓	✓	✓		✓
Caption Input	✓	✓		✓	✓	✓
ImageNet 1k	41.1	39.9	37.6	40.8	**41.2**	40.2
ImageNet Sketch	30.9	31.1	26.9	31.9	**31.9**	31.7
ImageNet V2	**34.1**	33.3	30.6	33.8	34.1	33.4
ImageNet-A	7.1	**7.5**	6.2	6.8	7.2	7.2
ImageNet-O	**48.9**	47.8	46.0	**48.9**	48.1	48.0
ImageNet-R	**47.6**	47.5	42.5	47.4	47.5	47.5
Caltech-101	81.7	80.4	78.9	80.8	80.7	**83.8**
CIFAR-10	89.8	88.2	83.8	88.1	88.4	**89.8**
CIFAR-100	63.8	63.6	59.2	65.2	65.0	**65.5**
CLEVR Counts	14.9	**26.2**	13.1	24.3	17.1	25.0
CLEVR Distance	**21.2**	18.6	16.4	15.9	15.9	15.8
SVHN	**26.8**	10.6	20.4	21.9	9.8	23.1
DTD	28.0	26.1	22.0	27.7	27.8	**28.7**
EuroSAT	35.9	**40.9**	22.5	31.4	36.5	32.6
KITTI distance	20.5	28.7	16.7	27.1	**34.2**	32.1
Oxford Flowers-102	38.8	35.8	39.3	39.3	**39.7**	36.3
Oxford-IIIT Pet	59.5	60.0	57.0	58.8	**61.3**	55.4
PatchCamelyon	57.5	54.7	56.8	52.3	**58.7**	53.1
RESISC45	31.0	34.8	28.7	**36.7**	33.9	34.5
FGVC Aircraft	3.3	3.2	3.3	2.6	**3.5**	3.4
Food-101	56.1	54.5	52.8	**56.5**	55.4	56.1
GTSRB	15.5	18.6	13.9	17.1	17.1	**19.7**
MNIST	29.8	22.8	23.4	29.5	26.1	**31.8**
ObjectNet	28.5	**28.7**	24.3	28.6	28.0	28.4
Pacal VOC 2007	63.8	70.2	54.7	67.6	69.1	**71.0**
Rendered SST2	50.2	50.1	**50.4**	49.9	49.2	49.7
Stanford Cars	45.3	44.1	**48.9**	45.7	48.2	42.5
STL-10	89.9	89.9	87.1	89.8	90.0	**90.2**
SUN-397	48.7	47.4	44.5	48.8	48.7	**49.7**
Country211	5.0	4.8	4.5	5.3	**5.3**	5.3
iWildCam	2.9	2.2	2.3	2.5	**3.5**	2.6
Camelyon17	57.0	65.8	**67.8**	53.1	66.0	55.8
FMoW	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
Dollar Street	**49.3**	46.1	47.1	48.2	48.7	47.4
GeoDE	73.0	70.5	66.2	**74.0**	**74.0**	68.8
Flickr30k	40.8	42.3	29.5	42.0	39.6	**44.6**
MSCOCO	25.3	18.0	17.2	26.2	24.2	**26.7**
WinoGAViL	43.1	37.7	36.9	41.6	**46.4**	40.5
Avg. over 38 datasets	39.6	39.3	36.3	39.7	**40.1**	39.9

Table 9. Comparison of the performance of different data refinement pipelines. Compared to other LVLMs, Qwen2VL demonstrates superior performance. Despite a tenfold difference in parameter size, Qwen2VL-7B with GPT-4o SFT still exhibits performance comparable to the 72B model. Additionally, the inclusion of captions significantly enhances dataset quality.

Method	Ours	DFN
DataComp scale	Large	Large
Dataset size	146.6M	146.6M
ImageNet 1k	**70.6**	68.7
ImageNet Sketch	**57.3**	54.9
ImageNet V2	**63.1**	60.0
ImageNet-A	**39.1**	29.9
ImageNet-O	43.0	**53.5**
ImageNet-R	**80.1**	75.4
Caltech-101	**93.1**	91.2
CIFAR-10	**96.2**	94.8
CIFAR-100	**81.0**	79.1
CLEVR Counts	**27.5**	14.7
CLEVR Distance	**22.2**	20.0
SVHN	45.9	**48.5**
DTD	**51.5**	46.9
EuroSAT	47.6	**49.9**
KITTI distance	**43.0**	24.9
Oxford Flowers-102	69.0	**71.0**
Oxford-IIIT Pet	**89.5**	88.7
PatchCamelyon	47.5	**51.0**
RESISC45	**60.6**	56.0
FGVC Aircraft	11.3	**13.2**
Food-101	**87.8**	86.2
GTSRB	**54.4**	44.2
MNIST	**77.7**	61.5
ObjectNet	**60.6**	55.0
Pacal VOC 2007	**78.8**	75.0
Rendered SST2	**51.7**	51.2
Stanford Cars	**85.3**	85.1
STL-10	**98.1**	96.0
SUN-397	**69.7**	67.2
Country211	**15.9**	13.5
iWildCam	**12.2**	10.0
Camelyon17	46.2	**63.1**
FMoW	**15.1**	10.9
Dollar Street	**61.3**	60.3
GeoDE	**88.7**	87.3
Flickr30k	**77.9**	68.2
MSCOCO	**52.2**	43.7
WinoGAViL	**52.8**	51.8
Avg. over 38 datasets	**58.6**	55.9

Table 10. Training on VLM-150M yields state-of-the-art CLIP models. We evaluate these models using the DataComp evaluation protocol. For detailed comparisons on specific datasets, we also provide the reproduced results for DFN. The symbol † indicates the results that we reproduced. Due to some broken links in the dataset, the amount of data used in our reproduction is slightly lower than that in the original paper.

References

[1] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks, 2023. 1

[2] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets, 2023. 2

[3] Zhengfeng Lai, Haotian Zhang, Bowen Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, Yinfei Yang, and Meng Cao. Veclip: Improving clip training via visual-enriched captions, 2024. 1, 2

[4] Xianhang Li, Zeyu Wang, and Cihang Xie. Clipa-v2: Scaling clip training with 81.1 1

[5] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. 1