Supplementary Material

Algorithm 1 The training pipeline of NCU.

```
Input: The CLIP model $\{f_v, f_t\}$ pre-trained on dataset $\mathbb{D}'$, fine-tuning dataset $\mathbb{D} \subseteq \mathbb{D}'$, $m$ learnable prompt tokens, margin
       parameters $[\alpha, \beta]$, forget ratio $P\%$, scaling factor $\lambda$, Sinkhorn regularization $\epsilon$, and balanced factor $\gamma$.
// Learning Hardest Negative Semantics
for $e = 1 : HN\_Epochs$ do
    for $n = 1 : num\_steps$ do
        Sample a batched samples $\mathbb{B} = \{(v_i, t_i)\}_{i=1}^{N}$ from $\mathbb{D}$
        // identifying the retained set
        $\mathcal{W} = \{w_i\}_{i=1}^{N}$, $w_P = \mathcal{Q}_P(\mathcal{W})$
        $\mathbb{D}_{RT} = \{(v_i, t_i) \mid w_i > w_P, \forall (v_i, t_i) \in \mathbb{B}\}$
        // updating
        Train the prompt and $f_t$ on $\mathbb{D}_{RT}$ by minimizing $\mathcal{L}^{HN}$
// Unlearning the Twin Noisy Correspondence
for $e = 1 : UL\_Epochs$ do
    for $n = 1 : num\_steps$ do
        Sample a batched samples $\mathbb{B} = \{(v_i, t_i)\}_{i=1}^{N}$ from $\mathbb{D}$
        // identifying the forget and retained set
        $\mathcal{W} = \{w_i\}_{i=1}^{N}$, $w_P = \mathcal{Q}_P(\mathcal{W})$
        $\mathbb{D}_{FG} = \{(v_i, t_i) \mid w_i \leq w_P, \forall (v_i, t_i) \in \mathbb{B}\}$,  $\mathbb{D}_{RT} = \{(v_i, t_i) \mid w_i > w_P, \forall (v_i, t_i) \in \mathbb{B}\}$
        Construct the mask matrix $\boldsymbol{M}$ by $\mathbb{D}_{FG}$ and $\mathbb{D}_{RT}$
        Calculate the hardest-negative guided alignment $\boldsymbol{T}$
        // updating
        Train the $f_v$ and $f_t$ on $\mathbb{B}$ by minimizing $\mathcal{L}^{UL}$
Output: The CLIP model $\{f_v, f_t\}$ with strong robustness.
```

Algorithm 2 Sinkhorn algorithm for Calculating Eq.(9).

```
Input: Cost matrix $\bar{\boldsymbol{C}} \in \mathbb{R}_+^{N\times(N+1)}$, mask matrix $\boldsymbol{M} \in
       \mathbb{R}_+^{N\times(N+1)}$, row distribution $\boldsymbol{\mu} = \frac{1}{N}\mathbb{1}_N$, column
       distribution $\bar{\boldsymbol{\nu}} = \frac{1}{N+1}\mathbb{1}_{N+1}$, Sinkhorn regularization parameter $\epsilon$, and max iterations $it_{max}$.
Initialize $\boldsymbol{K} = \boldsymbol{M} \odot e^{\frac{-\bar{\boldsymbol{C}}}{\epsilon}}$, $\boldsymbol{b} \leftarrow \mathbb{1}_{N+1}, it \leftarrow 0$
// Run Sinkhorn iterations
while $it \leq it_{max}$ and $\boldsymbol{a}$, $\boldsymbol{b}$ not convergence do
    $\boldsymbol{a} \leftarrow \frac{\boldsymbol{\mu}}{\boldsymbol{K}\boldsymbol{b}}$ // per-element division
    $\boldsymbol{b} \leftarrow \frac{\bar{\boldsymbol{\nu}}}{\boldsymbol{K}^\top \boldsymbol{a}}$
Output: OT Plan $\hat{\boldsymbol{\Gamma}}^* = \text{diag}(\boldsymbol{a})\boldsymbol{K}\text{diag}(\boldsymbol{b})$.
```

1. Training Pipeline

In this section, we summarize the training pseudo-code of NCU in Algorithm. 1.

2. Sinkhorn Solver for Masked OT

In this section, we detail the fast approximation for calculating the optimal transport plan. To avoid the high computational overhead of the exact linear programming solver, we adopt the entropy-regularized OT model that seeks an approximate solution of the OT plan by the fast Sinkhorn-Knopp algorithm. Algorithm. 2 presents the detailed Sinkhorn process. We could observe that the Sinkhorn's iteration involves only matrix multiplication and exponential operations, making it computationally efficient.

3. Dataset Details

3.1. Fine-tuning Datasets

We utilize three popular vision-language datasets at different scales and noise: Conceptual Captions 3M (CC3M) [18], Conceptual Captions 12M (CC12M) [2], and YFCC15M-R (provided by[9], an LLM-recaptioned subset from the original YFCC100M [20]). The details are as follows:

CC3M consists of 3.3 million image-caption pairs filtered from 5 billion webpage content, where image descriptions are sourced from the HTML alt-text attribute. As some image URLs in the original dataset have expired, we employ the version provided by Hugging Face[1], which comprises 2,905,954 pairs for training and 13,443 pairs for validation. CC3M is estimated to include a minimum of 3% false pos-

[1] https://huggingface.co/datasets/pixparse/cc3m-wds

itive pairs in which some unknown images and captions are mismatched or weakly matched.

CC12M is produced in a similar way to CC3M. It contains 12.4 million image-caption pairs that cover a broader range of topics and visual concepts from the real world. As some image URLs in the original dataset are broken, we employ the version provided by Hugging Face[2] including 10,968,539 training Paris.

YFCC15M-R is a subset provided by RWKV-CLIP [9] from the large-scale multimedia dataset YFCC100M [20], where the LLaMA3-8B model is used to synthesize diverse description. Consequently, YFCC15M-R is a cleaner dataset with fewer false positive pairs. we employ the version provided by Hugging Face[3] including 15,060,992 pairs (we drop the last batch samples).

3.2. Downstream Datasets

Dataset	Classes	Train Size	Test Size
Caltech101 [8]	102	3060	6085
CIFAR-10 [12]	10	50000	10000
CIFAR-100 [12]	100	50000	10000
DTD [4]	47	3760	1880
FGVC Aircraft [14]	100	6667	3333
RenderedSST2 [17]	2	6920	1821
Flowers102 [15]	102	2040	6149
Food101 [1]	101	75750	25250
GTSRB [19]	43	26640	12630
OxfordPets [16]	37	3680	3669
RESISC45 [3]	45	25200	6300
SUN397 [21]	397	19850	19850
EuroSAT [10]	10	10000	5000
StanfordCars [11]	196	8144	8041
STL10 [5]	10	1000	8000
ImageNet1K [6]	1000	1281130	5000

Table 1. Dataset sizes for downstream image classification tasks.

Dataset	Test Size of Images	Test Size of Captions
Flickr30K [22]	1000	5000
MSCOCO [13]	5000	25000

Table 2. Dataset sizes for downstream image-text retrieval tasks.

Image Classification. We evaluate the zero-shot transfer ability for image classification on ImageNet and 15 widely used downstream datasets. The specific testing information about the downstream classification datasets is presented in Tab. 1. Besides, we use 4 representative downstream datasets, *i.e.*, SUN397, OxfordPets, Food101, and ImageNet1K, to conduct the linear probing experiments. The detailed training size can also be found in Tab. 1.

Image-text Retrieval. We apply our approach to two standard image-text retrieval datasets, Flickr30K and MSCOCO, to evaluate the zero-shot retrieval performance. Both datasets have five corresponding text annotations for each image, and the detailed information is shown in Tab. 2.

4. Implementation Details

Source of Pre-trained CLIP. A key advantage of NCU is its ability to endow robustness to open-source pre-trained models, thus preventing the computational overhead of training from scratch. Following this point, we select pre-trained CLIP models (ViT-B/16 architecture) provided by the LaCLIP [7] that were pre-trained on CC3M and CC12M, and the model weights can be accessed in github[4]. For the pre-trained CLIP with ViT-B/32, we used our implementation since there were no publicly available model weights.

Config	Value
Batch size	2,048
Optimizer	AdamW
Learning Rate for HN	3e-4
Learning Rate for UL	5e-5
Epochs for HN	2
Epochs for UL	8
Total Epochs	10
Adam Beta	$\beta_1, \beta_2 = (0.9, 0.98)$
Adam Eps	1e-8
weight_decay	0.2
Ratio P (%)	10
Bound Margin	$\alpha, \beta = (-0.7, -0.2)$
Scaling factor λ	10
Balance factor γ	0.7
Sinkhorn regularization ϵ	0.03

Table 3. Hyper-parameters of NCU on CC3M.

Robust Fine-tuning. We employed the standard ViT-B/16 and ViT-B/32 as our visual encoders. Visual and textual features are projected into a shared 512-D space. Specifically, we conducted experiments using both ViT-B/16 and ViT-B/32 on CC3M and CC12M, and employed ViT-B/32 for experiments on YFCC15M-R. All experiments are conducted on 16 NVIDIA V100 GPUs. And the detailed fine-tuning settings for CC3M, CC12M, and

[2] https://huggingface.co/datasets/pixparse/cc12m-wds

[3] https://huggingface.co/datasets/Kaichengalex/YFCC15M

[4] https://github.com/LijieFan/LaCLIP

YFCC15M-R can be found in Tab. 3, Tab. 4, and Tab. 5, respectively.

Zero-shot Classification. We implement the identical prompt ensemble strategy as CLIP, where each class label is expanded to sentences using a collection of prompt templates, such as "*a tattoo of the [classname]*" or "*a photo of a nice [classname]*". The specific prompt templates for different downstream tasks can be found in [5].

Config	Value
Batch size	2,048
Optimizer	AdamW
Learning Rate for HN	3e-4
Learning Rate for UL	5e-5
Epochs for HN	2
Epochs for UL	8
Total Epochs	10
Adam Beta	$\beta_1, \beta_2 = (0.9, 0.98)$
Adam Eps	1e-8
weight_decay	0.2
Ratio P (%)	5
Bound Margin	$\alpha, \beta = (-0.7, -0.2)$
Scaling factor λ	10
Balance factor γ	0.7
Sinkhorn regularization ϵ	0.03

Table 4. Hyper-parameters of NCU on CC12M.

Config	Value
Batch size	2,048
Optimizer	AdamW
Learning Rate for HN	3e-4
Learning Rate for UL	5e-5
Epochs for HN	2
Epochs for UL	8
Total Epochs	10
Adam Beta	β_1, β_2 = (0.9, 0.98)
Adam Eps	1e-8
weight_decay	0.2
Ratio P (%)	1
Bound Margin	α, β = (-0.7, -0.2)
Scaling factor λ	10
Balance factor γ	0.3
Sinkhorn regularization ϵ	0.03

Table 5. Hyper-parameters of NCU on YFCC15M-R.

Linear Probing. Following the mainstream setting, we train a linear classifier using L-BFGS on features extracted from the frozen visual encoder. For all linear probing experiments, we set the batch size to 1024 and use the AdamW optimizer with 0.01 weight decay. The learning rate is initialized at 3e-4 and decreased with a cosine schedule. Classifiers are trained for 60, 60, 60, and 90 epochs on SUN397, OxfordPets, Food101, and ImageNet1K, respectively.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014. 2

[2] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021. 1

[3] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 2

[4] 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. 2

[5] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011. 2

[6] 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. 2

[7] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[8] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006. 2

[9] Tiancheng Gu, Kaicheng Yang, Xiang An, Ziyong Feng, Dongnan Liu, Weidong Cai, and Jiankang Deng. Rwkv-clip: A robust vision-language representation learner. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4799–4812, 2024. 1, 2

[10] 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. 2

[11] 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. 2

[5] `https://github.com/openai/CLIP/blob/main/data/prompts.md`

[12] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[14] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 2

[15] 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. 2

[16] 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. 2

[17] 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. 2

[18] 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, 2018. 1

[19] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE, 2011. 2

[20] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 1, 2

[21] 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. 2

[22] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014. 2