TinyCLIP: CLIP Distillation via Affinity Mimicking and Weight Inheritance
—— Supplementary Material ——

This supplementary material presents additional details of Sec. 4.1 and 4.3.

- **Distillation Temperature.** In Appendix A, we study the sensitivity analysis of the temperature parameter τ in *Affinity Mimicking* of Sec. 4.1.
- **Training Settings.** In Appendix B, we provide the detailed training settings for TinyCLIP in Sec. 4.1.
- **Architectures in Automatic Inheritance.** In Appendix C, we expound the architectures of TinyCLIP before and after automatic weight inheritance.
- **Additional Analysis.** In Appendix D, we further investigate different interaction modes in *Impact of affinity mimicking*, as well as the efficacy of *weight inheritance* on single modality compression in Sec. 4.3.

A. Distillation Temperature

As shown in Fig. 7, when $\frac{1}{\tau}$ is 50, it obtains the best accuracy. If $\frac{1}{\tau}$ is lower than 30, the probability distribution becomes uniform, which prevents the model from converging. Furthermore, when $\frac{1}{\tau}$ exceeds 100, the probability distribution approaches a hard label of 0 or 1, which results in a lack of transfer of relationship knowledge from the teacher model to the student model. Therefore, we set the temperature parameter to $\frac{1}{50}$ by default.

B. Training Settings

Tab. 8 presents an overview of hyperparameters used for training on LAION-400M [9] and YFCC-15M [11] datasets. The compression process is divided into three stages on LAION-400M [9] or two stages on YFCC-15M [11], in consistency with Sec. 4.1 of the main manuscript. We also combine LAION-400M [9] and YFCC-15M [11] as a training set in Sec. 4.2, named LAION+YFCC-400M, where the number of pairs is 400M, and each batch consists of 2:1 pairs sampled from the two datasets. Its training setting is the same as that on LAION-400M [9].

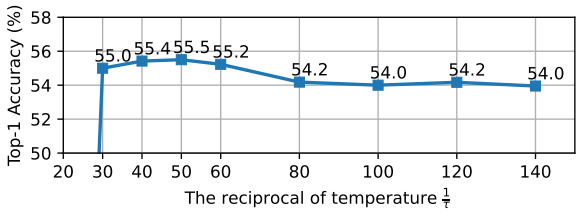

Figure 7. Ablation study on the distillation temperature parameter. The student model TinyCLIP ViT-40M/32 is inherited and distilled on LAION-400M [9] for 1 epoch. The zero-shot accuracy on ImageNet-1K [1] is reported.

Hyper-parameter	LAION-400M	YFCC-15M
Batchsize	32,768	4,096
Optimizer	AdamW [6, 4]	
Optimizer Momentum	$\beta_1 = 0.9, \beta_2 = 0.98$	
Base Learning Rate	10^{-4}	
Weight Decay	0.2	
Learning Rate Schedule	Cosine decay [5]	
Warmup (steps)	2,000	
Gradient Clipping Norm	5	
The Reciprocal of Temperature	50	
Image Resolution	224×224	
Image Augmentation	RandomResizedCrop	
Tokenizer	Byte Pair Encoding [10]	
Vocabulary size	49,408	
Max Sequence Length	77	

Table 8. Training settings on LAION-400M [9] and YFCC-15M [11] datasets.

C. Architectures in Automatic Inheritance

In this section, we utilized CLIP ViT-B-32 [7, 3] as an example to illustrate the architecture after 50% automatic weight inheritance.

	Vision Transformer			Text Transformer			# params (M)		
Model	width	heads	inter	width	heads	inter	vision	text	total
CLIP ViT-B/32	768	12	3072	512	8	2048	88	38	126
TinyCLIP ViT-45M/32	537	9.3	2044.7	508	5	749.4	45	18	63

Table 9. Encoder specifics. Since TinyCLIP with automatic inheritance has different channels per layer, we report the average number of all 12 layers. Let width, heads and inter denote embedding dimension, MHA heads and FFN channels, respectively.

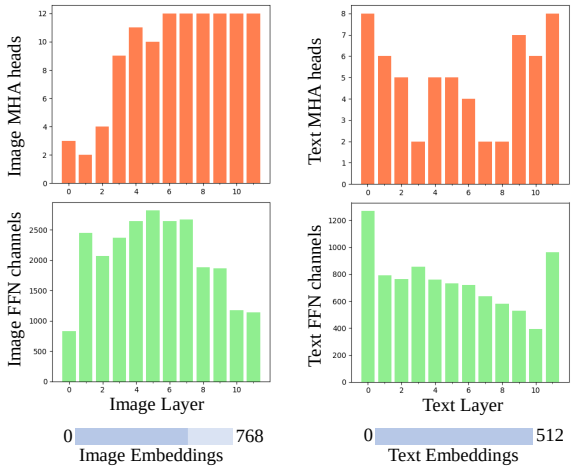

Figure 8. The number of remaining MHA heads, FFN channels and embeddings of CLIP image and text encoder after 50% automatic weight inheritance. The orange bars represent MHA heads, the green bars denote FFN channels, the blue bars refer to embedding channels.

CLIP ViT-B/32 [7] adopts ViT-B/32 [2] (L_c=12, N_h=12, d=768, p=32) as the image encoder and Transformer [12] (L_c=12, N_h=8, d=512) as the text encoder, where L_c represents the layers, N_h denotes the MHA (Multi-Head Attention) heads, d refers to the embedding dimension and p is the patch size. Consistent with automatic inheritance described in Sec. 3.2, we select important weights by inserting masks.

As shown in Fig. 8 and Tab. 9, we present the structure of TinyCLIP ViT-45M/32 after automatic weight inheritance. For the image encoder, MHA heads and FFN (Feed-Forward Network) channels are kept more than text encoder, and the discarded parts are mainly concentrated in the first few or the last layers. Besides, embeddings of image encoder are compressed from 768 to 537 channels. For the text encoder, MHA heads and FFN channels in transformer layer are retained fewer, and the discarded parts are mainly concentrated in the middle layers, while embeddings are almost all kept. This reflects the redundancy difference between image branch and text branch, where image encoder has more redundancy in width (embeddings) and text encoder has more redundancy in depth (transformer layers).

We speculate that this redundancy difference is caused by the information density of images and text. Image data has low information density, thus more MHA heads and FFN channels are needed to handle details and high-level features when encoding images, so as to enhance the expressiveness of the model. In contrast, text data has high information density, so the embeddings are kept to facilitate the model's learning of the semantic information.

Loss	Similarity		IN-1K	Flickr30k		MSCOCO	
	single-modal	cross-modal	top1 acc(%)	I→T R@1	T→I R@1	I→T R@1	T→I R@1
Contrastive loss							
$\mathcal{L}_0$	-	-	53.4	71.8	53.1	44.4	28.7
Affinity mimicking							
$\mathcal{L}_0 + \mathcal{L}_1$	0.74	0.32	55.1	71.8	54.0	46.4	29.6
$\mathcal{L}_1$ (or $\mathcal{L}_{distill}$)	0.77	0.32	55.5	74.0	55.0	46.7	30.9
Cross modalities							
$\mathcal{L}_2 + \mathcal{L}_3$	0.88	0.37	55.3	72.4	54.8	46.5	30.0
$\mathcal{L}_1 + (\mathcal{L}_2 + \mathcal{L}_3)$	0.88	0.37	56.2	72.9	55.1	47.1	30.5
Single modality							
$\mathcal{L}_4 + \mathcal{L}_5$	0.57	0.50	19.2	38.2	23.4	20.5	10.0
$\mathcal{L}_1 + (\mathcal{L}_4 + \mathcal{L}_5)$	0.69	0.35	55.2	73.1	55.5	46.9	30.1

Table 10. Ablation study on different interactions. The model OpenCLIP ViT-B/32 [3] pre-trained on LAION-2B [8] is the teacher, inherited to the student model TinyCLIP ViT-40M/32 on LAION-400M [9] for 1 epoch. The average similarity of student's embedding feature and teacher's one is reported.

D. Additional Analysis

In this section, we provide the detail of different interactions of distillation losses in Sec. 4.3, then study the effect of single modality compression to verify the redundancy differences between image and text encoder.

D.1. Different Interaction Modes

We provide more metrics for Tab. 2 in the manuscript. As shown in Tab. 10, affinity mimicking $\mathcal{L}_1$ outperforms contrastive loss $\mathcal{L}_0$ by 2.1% zero-shot top-1 accuracy on ImageNet [1]. Using $\mathcal{L}_1$ alone is better than the combination of $\mathcal{L}_0$ and $\mathcal{L}_1$. Furthermore, when combining with $\mathcal{L}_1$, cross modalities $\mathcal{L}_2$ and $\mathcal{L}_3$ improves the zero-shot accuracy on ImageNet by 2.8% compared to $\mathcal{L}_0$. We observe that cross modalities interaction brings a high single-modal similarity of 0.88, enabling students' embedding space to align with the teacher's one. However, single modality interaction do not work. We conjecture that the interaction on the same modality brings little information, since the teacher model is only trained by the interaction of different modalities.

D.2. Single Modality Compression Analysis

We evaluate the impact of compression on single modality with manual weight inheritance. As shown in Tab. 11, when removing the last 4 layers in the text encoder, the 0-epoch accuracy on ImageNet [1] is 24.0% (#1), while for the image encoder, the 0-epoch accuracy is 0% (#5). It indicates that the layer-wise redundancy of the text encoder is larger than that of the image encoder.

Moreover, when training student models for 1 epochs, weight inheritance brings 3.6% accuracy for text encoder (#3 *vs.* #2) and 9.8% accuracy for image encoder (#7 *vs.* #6). It demonstrates weight inheritance brings a good initialization for student models.

To further investigate the impact of compression, we train #4 for the short 5 epochs, which achieves 64.2% ac-

#	Image depth -width	Text depth -width	#Params image+text (M)	Ep.	IN-1K top1 acc (%)	MSCOCO I→T R@1	MSCOCO T→I R@1
0	12-768	12-512	88+38	32	65.7	56.9	39.3
Weight inheritance for text encoder							
1	12-768 ❄	8-512 ♲	88+26	0	24.0	11.1	7.1
2	12-768 ❄	8-512	88+26	1	59.1	49.0	32.9
3	12-768 ❄	8-512 ♲	88+26	1	62.7	53.2	35.8
4	12-768 ❄	8-512 ♲	88+26	5	64.2	53.5	36.4
Weight inheritance for image encoder							
5	8-768 ♲	12-512 ❄	60+38	0	0.0	0.0	1.7
6	8-768	12-512 ❄	60+38	1	41.5	36.1	21.1
7	8-768 ♲	12-512 ❄	60+38	1	51.3	46.3	29.1
8	12-640 ♲	12-512 ❄	61+38	1	58.2	51.2	33.9

Table 11. Ablation study on compression for single modality. The image and text encoders are both transformers, where ❄ represents frozen weights, and ♲ represents weight inheritance. The weights are inherited from OpenCLIP ViT-B/32 [3], pre-trained on LAION-2B [8]. All models are trained on LAION-400M [9] without distillation.

curacy with 12M (31.6%) parameters reduction on text encoder. Besides, we reduce the width of image encoder (#8), which surpasses reducing the depth of image encoder (#7) by significant 6.9% accuracy. It indicates that compressing along the width dimension is better than the depth dimension for image encoder.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 2

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2

[3] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. 1, 2, 3

[4] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015. 1

[5] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 1

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018. 1

[7] 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 *ICML*, 2021. 1, 2

[8] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 2, 3

[9] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1, 2, 3

[10] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *ACL*, pages 1715–1725, 2016. 1

[11] 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*, 2016. 1

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2