FuseTeacher: Modality-fused Encoders are Strong Vision Supervisors

Chen-Wei Xie[1], Siyang Sun[1], Liming Zhao[1],
Pandeng Li[2], Shuailei Ma[3], and Yun Zheng[1]

[1] Alibaba Group
{eniac.xcw,siyang.ssy,lingchen.zlm}@alibaba-inc.com
[2] University of Science and Technology of China
[3] Northeastern University

Appendix

This appendix is structured as follows.

- Sec. A provides the implementation details of the fusion encoder.
- Sec. B describes the collection process for the Union23M/Union65M datasets.
- Sec. C presents additional ablation experiments.
- Sec. D offers further analysis to substantiate the superior discriminative power of the fusion representation compared to the image representation.
- Sec. E applies the FuseTeacher to video tasks.
- Sec. F discusses the training efficiency of the proposed FuseTeacher model.
- Sec. G provides an overview of the downstream datasets we used.
- Sec. H discusses the relationship between the FuseTeacher and previous method RA-CLIP [35].

A Fusion Encoder

The architecture of the fusion encoder is built upon transformer blocks that incorporate cross-attention mechanisms. Fig. 1 illustrates two different ways to implement the fusion encoder. The *standard* implementation in Fig. 1 (a) is similar to the multi-modal encoder used in previous works, such as ALBEF [21] and BLIP [20]. It begins with a self-attention operation, then uses the resulting output as a query, along with the embeddings of the corresponding text tokens as key and value, to perform cross-attention. The goal of the cross-attention is to enhance the image embeddings with the text embeddings. Finally, a standard multi-layer perceptron is used, and the embedding for the image [CLS] token is extracted as the fused representation.

Furthermore, to reduce computation for the fusion encoder, we also introduce a *fast* implementation. As shown in Fig. 1 (b), it takes a single image [CLS] embedding as the query and performs cross-attention with image patch embeddings and text tokens embeddings. Because the query for the two cross-attentions and the input for the multi-layer perception are single embeddings, it is extremely fast.

The performance and training efficiency of these two implementations are reported in Sec. F. Both implementations achieve similar performance. Our main purpose is to validate the effectiveness of modality-fused supervision, rather than designing a better fusion encoder module. Therefore, in the main paper, we adopt the widely-used *standard* implementation by default.

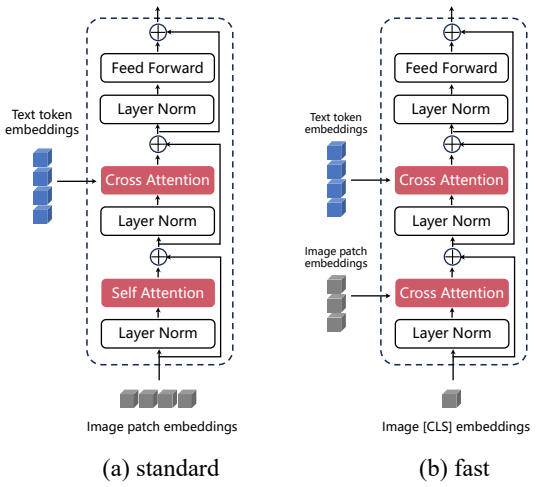

Fig. 1: Implementation details of the fusion encoder.

B Union23M and Union65M

The YFCC15M [33] dataset is widely used in recent contrastive language-image pre-training (CLIP) works. However, it is observed that models trained on the YFCC15M dataset exhibit markedly inferior performance related to the OpenAI CLIP models [28]. This raises doubts about the effectiveness of existing algorithms when compared to a strong baseline. However, training models on hundreds of millions of image-text pairs, as done by OpenAI, requires extensive GPU resources that are beyond the budget of most labs and companies. To address this challenge, we incorporate approaches from recent advancements including CiT [36], DataComp [13], and DINO v2 [25]. By employing selective sampling techniques on extensive datasets such as LAION400M [31], LAION2B [30], COYO700M [3], YFCC15M [33], CC12M [5], CC3M [32], and SBU [26], we have curated two novel datasets, termed Union23M and Union65M.

Specifically, we sample the datasets based on two key aspects:

– **Visual Curation.** We extract embeddings of all images using the DINO v2 *ViT-S/14* model [25], and cluster them into 100,000 groups. We then calculate the embeddings of ImageNet training images and determine their

nearest clusters. We retain the images that belong to the same clusters as the ImageNet training images. This process is inspired by DINO v2 [25] and DataComp [13].
– **Textual Curation.** CiT proposes to sample image-text pairs based on their textual similarity with ImageNet categories. Instead of using CLIP text similarity for this, we employ keyword matching in our approach, retaining pairs that contain ImageNet categories in their text descriptions.

Following the aforementioned methodology, we extract 23 million image-text pairs from the union set of LAION400M, COYO700M, YFCC15M, CC12M, CC3M, and SBU, resulting in a new dataset Union23M. Additionally, we use LLaVA to create supplementary text description for each image. Following the same method, we sample 65 million image-text pairs from the union set of LAION2B, COYO700M, YFCC15M, CC12M, CC3M, and SBU, resulting in a larger dataset named Union65M. These two datasets will be released. We believe it will accelerate the research on Vision-Language Pre-training, especially for the laboratories and companies who cannot afford to training on hundreds of millions of image-text pairs.

The CLIP models trained on these datasets are strong than the baseline models of some previous works [10, 19, 23, 37]. Improving on this strong baseline is more challenging. However, Table 6 in the main paper demonstrates that the proposed FuseTeacher still outperforms this strong baseline, which further validates the effectiveness of it.

C More Ablation Experiments

C.1 The Number of Clustering Centroids

In Sec. 3.3 of the main paper, we cluster the image-text pairs into $K = 4096$ groups using a soft assignment strategy. In this section, we extend our experiments by setting K to 128,000. The top-1 accuracy on the zero-shot ImageNet classification task is 43.4%, which is on par with the 43.5% achieved using $K = 4096$. These experimental findings suggest that $K = 4096$ is an adequate choice for datasets like YFCC15M, and that the model's performance is not sensitive to K when K is large enough.

C.2 The Number of Fusion Blocks

We implement the fusion encoder with two transformer blocks in the main paper. In this section, we conduct experiments to determine whether a larger fusion encoder can yield further improvements. Specifically, we implement a larger fusion encoder with six transformer blocks. This results in a marginally better (+0.1%) multi-modal classification performance on the ImageNet-MM dataset, and maintains the same image classification result (43.5%) on the ImageNet 1K. Given that the image and text have already been processed by their respective encoders, a lightweight fusion encoder appears to be sufficient for their integration.

Table 1: Training efficiency of different methods.

	FuseTeacher	MT loss	zero-shot	k-NN	linear-probe	train time
CLIP [28]	-	✗	41.1	59.1	66.5	1.0×
SLIP [23]	-	✗	42.2	57.4	66.5	**2.1**×
MaskCLIP [10]	-	✗	43.1	61.3	68.2	1.4×
CLIP [28]	-	✓	44.3	61.6	68.4	1.4×
CLIP	standard	✗	43.5	61.5	68.4	1.5×
CLIP	standard	✓	45.0	**63.0**	69.6	1.5×
CLIP	fast	✓	**45.1**	**63.0**	**69.7**	1.4×

D More Analysis on Fusion Representation

In the Sec. 4.2 of the main paper, we demonstrate that the fusion encoder outperforms the image encoder on the ImageNet-MM, COCO-MM, and Flickr-MM datasets. In this section, we also show that the fusion encoder also excels over the image encoder on the training data during pre-training.

As shown in Fig. 2, loss image-text denotes the contrastive loss between image embeddings and text embeddings, i.e., $\mathcal{L}_{v2t} + \mathcal{L}_{t2v}$. Loss fusion-text denotes the contrastive loss between fusion embeddings and text embeddings, i.e., $\mathcal{L}_{f2t} + \mathcal{L}_{t2f}$. Loss fusion-text is markedly lower than loss image-text, signifying superior performance of the fusion embeddings during training. Coupled with our experiments on ImageNet-MM, COCO-MM, and Flickr-MM in Sec. 4.2, we conclude that fusion embeddings outperform image embeddings in both training and testing stages, establishing them as an effective teacher.

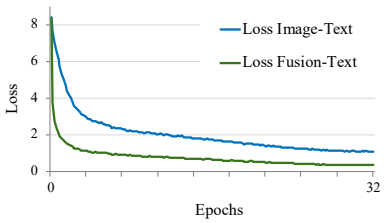

Fig. 2: The contrastive losses between image-text pairs and fusion-text pairs.

E Experiments on Video

We also conducted experiments on video dataset. Since pre-training on large-scale video datasets requires much time, we fine-tune pre-trained video-text models on small dataset YouCook2. We apply FuseTeacher to VAST [6], utilizing VAST's video+audio+subtitles fusion feature as modality-fused supervision. Since YouCook2 is small, we cannot significantly update the pre-trained model. Instead, we fine-tune it with a small lr 3e-5. Nonetheless, in text-to-video retrieval task (no audio/subtitles at inference), FuseTeacher achieves R@1, R@5, R@10 scores of 21.5/48.8/61.1, surpassing the fine-tuned VAST's 20.4/47.6/59.6.

Table 2: Details of the downstream datasets.

	Dataset	# Classes	# Train	# Val/Test
Zero-shot classification & linear probing	ImageNet [9]	1,000	1,281,167	50,000
	Pets [27]	37	3,680	3,669
	CIFAR10 [18]	10	50,000	10,000
	CIFAR100 [18]	100	50,000	10,000
	SUN397 [34]	397	19,850	19,850
	Food101 [2]	101	75,750	25,250
	Flowers [24]	102	2,040	6,149
	Cars [17]	196	8,144	8,041
	Caltech101 [12]	102	3,060	6,085
	Aircraft [22]	100	6,667	3,333
	DTD [8]	47	3,760	1,880
Image–text retrieval	Flickr30K [38]	-	-	1,000
	MS COCO [7]	-	-	5,000
ImageNet variations	ImageNet-R [15]	200	-	30,000
	ImageNet-Sketch [14]	1,000	-	50,000
	ImageNetV2 [29]	1,000	-	30,000
	ImageNet-A [16]	200	-	7,500
	ObjectNet [1]	113	-	18,574

F Training Efficiency

The main computational overhead of FuseTeacher compared to CLIP is processing an additional text input during training. However, it's important to note that FuseTeacher's computational cost is comparable to, or even less than, that of recent CLIP variants. For example, SLIP and UniCLIP require 3× more image processing during training. DeCLIP processes 2× images and 2× texts. MaskCLIP adds a momentum encoder and takes an extra masked image input.

Although processing additional text input for FuseTeacher is inevitable, we can introduce a multi-text contrastive loss (MT loss) to fully utilize the additional text computation. Specifically, we modify the image-to-text contrastive loss in Eq. 4 in the main paper as follows:

$$\mathcal{L}_{v2t} = \frac{1}{2} \sum_{m \in \{a,b\}} -log(\frac{exp(\sigma(\mathbf{v}_i, \mathbf{t}_i^m)/\tau_1)}{\sum_{j=1}^{N} exp(\sigma(\mathbf{v}_i, \mathbf{t}_j^m)/\tau_1)}), \tag{1}$$

where each image corresponds to two text embeddings $\mathbf{t}_i^a$ and $\mathbf{t}_i^b$. This formula is similar to the multi-text contrastive loss used in LaCLIP [11]. Our experiments demonstrate that the multi-text contrastive loss can provide additional improvements.

Table 1 summarizes the performance and training time of different methods. We find that FuseTeacher brings consistent improvements over different baseline methods with minimal or no additional training cost.

G Details of Downstream Datasets

Table 2 demonstrates the details of each dataset we used for downstream tasks.

H Relation between FuseTeacher and RA-CLIP

Previous method RA-CLIP [35] utilizes DINO-S/8 [4] to retrieve relevant image-text pairs, and use the text descriptions to enhance the image embedding. Although effective, it introduces more computation than vanilla CLIP during inference. Specifically, RA-CLIP involves extracting DINO-S/8 features and performing retrieval operations. In contrast, FuseTeacher only employs additional text descriptions during training. Once trained, FuseTeacher can be deployed in a manner similar to vanilla CLIP, making it more efficient.

References

1. Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., Katz, B.: ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In: Adv. Neural Inform. Process. Syst. pp. 9448–9458 (2019)
2. Bossard, L., Guillaumin, M., Gool, L.V.: Food-101–mining discriminative components with random forests. In: Eur. Conf. Comput. Vis. pp. 446–461 (2014)
3. Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., Kim, S.: COYO-700M: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset` (2022)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision Transformers. In: Int. Conf. Comput. Vis. pp. 9650–9660 (2021)
5. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3558–3568 (2021)
6. Chen, S., Li, H., Wang, Q., Zhao, Z., Sun, M., Zhu, X., Liu, J.: Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. Adv. Neural Inform. Process. Syst. **36** (2024)
7. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft COCO captions: Data collection and evaluation server (2015)
8. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3606–3613 (2014)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 248–255 (2009)

10. Dong, X., Bao, J., Zheng, Y., Zhang, T., Chen, D., Yang, H., Zeng, M., Zhang, W., Yuan, L., Chen, D., et al.: MaskCLIP: Masked self-distillation advances contrastive language-image pretraining. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10995–11005 (2023)
11. Fan, L., Krishnan, D., Isola, P., Katabi, D., Tian, Y.: Improving CLIP training with language rewrites. In: Adv. Neural Inform. Process. Syst. (2023)
12. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. pp. 178–178 (2004)
13. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., Orgad, E., Entezari, R., Daras, G., Pratt, S.M., Ramanujan, V., Bitton, Y., Marathe, K., Mussmann, S., Vencu, R., Cherti, M., Krishna, R., Koh, P.W., Saukh, O., Ratner, A., Song, S., Hajishirzi, H., Farhadi, A., Beaumont, R., Oh, S., Dimakis, A., Jitsev, J., Carmon, Y., Shankar, V., Schmidt, L.: DataComp: In search of the next generation of multimodal datasets (2023)
14. Gao, S., Li, Z., Yang, M., Cheng, M., Han, J., Torr, P.H.S.: Large-scale unsupervised semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **45**(6), 7457–7476 (2023)
15. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., Gilmer, J.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Int. Conf. Comput. Vis. pp. 8320–8329 (2021)
16. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 15262–15271 (2021)
17. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Int. Conf. Comput. Vis. Worksh. pp. 554–561 (2013)
18. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
19. Lee, J., Kim, J., Shon, H., Kim, B., Kim, S.H., Lee, H., Kim, J.: UniCLIP: Unified framework for contrastive language-image pre-training. In: Adv. Neural Inform. Process. Syst. pp. 1008–1019 (2022)
20. Li, J., Li, D., Xiong, C., Hoi, S.C.H.: BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Int. Conf. Mach. Learn. pp. 12888–12900 (2022)
21. Li, J., Selvaraju, R.R., Gotmare, A.D., Joty, S.R., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. In: Adv. Neural Inform. Process. Syst. (2021)
22. Maji, S., Rahtu, E., Kannala, J., Blaschko, M.B., Vedaldi, A.: Fine-grained visual classification of aircraft (2013)
23. Mu, N., Kirillov, A., Wagner, D.A., Xie, S.: SLIP: self-supervision meets language-image pre-training. In: Eur. Conf. Comput. Vis. vol. 13686, pp. 529–544 (2022)
24. Nilsback, M., Zisserman, A.: Automated flower classification over a large number of classes. In: ICVGIP. pp. 722–729 (2008)
25. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision (2023)
26. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. Adv. Neural Inform. Process. Syst. **24** (2011)

27. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3498–3505 (2012)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn. pp. 8748–8763. PMLR (2021)
29. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do ImageNet classifiers generalize to ImageNet? In: Int. Conf. Mach. Learn. vol. 97, pp. 5389–5400 (2019)
30. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S.R., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5b: An open large-scale dataset for training next generation image-text models. In: Adv. Neural Inform. Process. Syst. (2022)
31. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs (2021)
32. Sharma, P., Ding, N., Goodman, S., Soricut, R.: 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. pp. 2556–2565 (2018)
33. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: YFCC100M: the new data in multimedia research. Commun. ACM **59**(2), 64–73 (2016)
34. Xiao, J., Ehinger, K.A., Hays, J., Torralba, A., Oliva, A.: SUN database: Exploring a large collection of scene categories. Int. J. Comput. Vis. **119**(1), 3–22 (2016)
35. Xie, C.W., Sun, S., Xiong, X., Zheng, Y., Zhao, D., Zhou, J.: RA-CLIP: Retrieval augmented contrastive language-image pre-training. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19265–19274 (2023)
36. Xu, H., Xie, S., Huang, P.Y., Yu, L., Howes, R., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C.: CiT: Curation in training for effective vision-language data. In: Int. Conf. Comput. Vis. pp. 15180–15189 (2023)
37. You, H., Zhou, L., Xiao, B., Codella, N., Cheng, Y., Xu, R., Chang, S.F., Yuan, L.: Learning visual representation from modality-shared contrastive language-image pre-training. In: Eur. Conf. Comput. Vis. pp. 69–87. Springer (2022)
38. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Trans. Assoc. Comput. Linguistics **2**, 67–78 (2014)