Supplementary Materials for Long-CLIP: Unlocking the Long-Text Capability of CLIP

Beichen Zhang^{§1,2}, Pan Zhang¹, Xiaoyi Dong^{1,3*}, Yuhang Zang¹, and Jiaqi Wang^{1*}

¹Shanghai AI Laboratory ²Shanghai Jiao Tong University ³The Chinese University of Hong Kong zhangbeichen@sjtu.edu.cn, {zhangpan, dongxiaoyi, zangyuhang, wangjiaqi}@pjlab.org.cn https://github.com/beichenzbc/Long-CLIP

1 Urban-1k Dataset

Urban-1k is a scaling-up version of Urban-200 dataset in the paper. It contains 1k urban images and their corresponding captions generated by GPT-4V. Each caption contains about 107 words on average. The process for building Urban-1k dataset is exactly the same as Urban-200. It has been released at https://huggingface.co/datasets/BeichenZhang/Urban1k.

	Model	Urban-1k I2T R@1	Urban-1k T2I R@1
B/16	CLIP	68.1	53.6
	Long-CLIP(Ours)	78.9	79.5
L/14	CLIP	68.7	52.8
	Long-CLIP(Ours)	82.7	86.1

Table 1: The result on Urban-1k dataset. Best result is in **bold**.

2 Insufficient Long-text Ability for CLIP-based Models

Apart from aligning the whole image with the whole sentence, a few recent works (e.g., PTP-BLIP [11] and X-VLM [12]) have tried to align some text phrases in the whole caption with their corresponding image regions (patches) through contrastive learning and claim to improve the ability to capture fine-grained information. We evaluate these models on our urban dataset with long captions, and the results are shown in Tab. 2. We observe that the long-text performance of these methods still has room for improvement since their training datasets predominantly consist of short texts (e.g., Conceptual-12M [1], SBU [8], Visual Genome [4] and COCO [7]). By contrast, our Long-CLIP (bottom row) boosts the performance of these baselines by a large margin.

^{*} Corresponding author. § Work done during an internship in Shanghai AI Laboratory.

2 B. Zhang et al.

Table 2: The long-text performance of recent advances attempting to improve the fine-grained capability. X-VLM(16M) indicates the version of X-VLM base model with 16M training data.

Model	Image-to-Text	Text-to-Image	Training Data
CLIP [10]	46.5	46.0	WIT
X-VLM(16M) [12]	53.0	44.5	CC12M+CC3M+SBU+COCO+VG
X2-VLM [13]	45.0	45.0	CC3M+SBU+COCO+VG+CC12M+LAION
PTP-BLIP [11]	45.5	39.5	$\rm CC3M+SBU+COCO+VG$
Long-CLIP(Ours)	81.5	81.5	WIT+ShareGPT4V

3 Long-CLIP with SDXL

After the first submission, we further use our Long-CLIP model in Stable-Diffusion-XL [9] in a plug-and-play manner. Specifically, we replace the CLIP-L text encoder with our Long-CLIP-L, and only apply knowledge-preserved stretching (KPS) on Open-CLIP bigG text encoder due to heavy training cost.

The results are shown in Fig. 1, which shows our model can help SDXL break the 77 token limit with little reduction in the image quality.

4 Generalizability of Proposed Strategy

We apply our strategy, namely knowledge-preserved stretching and primary component matching, to fine-tune DeCLIP ViT-B/32 [6], the result also shows a consistent improvement.

Table 3: The performance of DeCLIP model and Long-DeCLIP model after finetuning. 'Avg Cls' means the average accuracy among 5 classification datasets in the main paper. Retrieval tasks are reported in 'T2I/I2T' format with R@5 of short-caption (COCO/Flickr) and R@1 of long-caption (Urban/ShareGPT4V).

Model	Avg Cls	COCO	Flickr	Urban	ShareGPT4V
DeCLIP Long-DeCLIP	63.6 63.9	$\begin{array}{c c} 60.5/45.8 \\ 61.1/50.7 \end{array}$	36.2/25.3 36.8/33.8	$\begin{array}{c} 28.0/25.5 \\ 54.0/51.0 \end{array}$	$\begin{array}{c c} 63.1/60.7 \\ 84.1/84.3 \end{array}$

5 Detailed Experiment Setting

5.1 Long Text Fine-tuning

We fine-tune the CLIP model on ShareGPT4V [2] dataset, which contains about 1M (image, long caption) pairs in total. Detailed hyper-parameter settings in long-text fine-tuning are listed in Tab. 4.

The serene lake surface resembled a flawless mirror, reflecting the soft blue sky and the surrounding greenery. A gentle breeze played across its expanse, ruffling the surface into delicate ripples that gradually spread out, disappearing into the distance. Along the shore, weeping willows swayed gracefully in the light breeze, their long branches dipping into the water, creating a soothing sound as // they gently brushed against the surface. In the midst of this serene scene, a pure white swan floated gracefully on the lake. Its elegant neck curved into a graceful arc, giving it an air of dignity.



CLIP

Long-CLIP

The painting captures a serene moment in nature. At the center, a calm lake reflects the sky, its surface rippled only by the gentlest of breezes. The sky above is a brilliant mix of blues and whites, with fluffy clouds drifting leisurely across. On the banks of the lake, tall trees stand gracefully, their leaves rustling in the wind. // In the foreground, an old man sits on a rock, seemingly lost in deep thought or meditation. The soft light of the setting bathes the entire scene in a warm glow, creating a sense of peace and tranquility. The colors are muted yet vibrant, and the details are captured with precision, giving the painting a sense of realism while still retaining a dreamlike quality.



CLIP

Long-CLIP

A captivating scene unfolds in this painting. The backdrop is a vibrant sunset, its hues ranging from vibrant orange to deep purple, painting the sky in a stunning array of colors. Beneath this canopy of color, a lush green meadow stretches out, dotted with wildflowers of various hues. In the foreground, a solitary tree stands tall, its branches reaching towards the sky as if embracing // the setting sun. Its leaves rustle gently in the evening breeze, creating a serene soundtrack. A bride stood on the grass in a pristine white wedding dre holding a bouquet in her hands, seemingly unaware of the breathtaking beauty around her, adding a touch of humanity to this serene natural scene.



CLIP

Long-CLIP

Fig. 1: Our Long-CLIP model can help SDXL break through the 77-token limit (marked in red \)) and capture the originally truncated attribute (marked in orange) with little reduction in the image quality.

5.2 **Zero-shot Classification**

We follow the setup of CLIP [10]. For zero-shot classification tasks like ImageNet [3] and CIFAR-100 [5], we use the 80 pre-defined prompts used in CLIP. We compute the embedding of each class by averaging over the embeddings of the 80 prompts. Then we L2-normalize them. For a given image in a classifica-

4 B. Zhang et al.

Table 4. Detailed in per-parameters in long-text inte-tun	Table 4:	Detailed	hyper-r	oarameters	in	long-text	fine-tunir
--	----------	----------	---------	------------	----	-----------	------------

Hyper-Parameter	Value
Batch size	1024
Training Epochs	1
Warm-up iterations	200
Weight decay	1e-2
Learning Rate	1e-4
AdamW β_1	0.9
AdamW β_2	0.999
Adam W ϵ	1e-8

tion dataset, we classify it as the class that has the largest cosine similarity with the image embedding. We use top-1 accuracy as an evaluation metric.

5.3 Retrieval

In text-image retrieval tasks, we calculate text-image scores by measuring the cosine similarity between the L2-normalized images and text embedding. We then rank the top-K images for each text caption, as well as the top-K text captions for each image. We use Recall@K as an evaluation metric where K can be 1, 5 and 10, which is a common setting for retrieval tasks.

6 More Examples on Retrieval and Image Generation

Our Long-CLIP model can capture detailed information in both image and text modalities. Therefore, Long-CLIP can distinguish similar images and texts and improve retrieval accuracy. Fig. 2 demonstrates some examples where CLIP fails but our Long-CLIP model can successfully retrieve. Moreover, as shown in Fig. 3, Long-CLIP can also enhance image generation tasks by covering more details in the text prompt compare to the CLIP baseline.

Text-Image Retrieval

The image captures a serene coastal town, nestled on a hillside. The town is characterized by a series of white houses, each topped with a vibrant red roof. These houses are scattered across the hillside, their white facades contrasting beautifully with the red roofs. The hillside itself is adorned with lush green trees and bushes, adding a touch of nature to the urban landscape. The town overlooks a deep blue body of water, which is visible in the foreground of the image. The water's surface is calm, reflecting the clear blue sky above. A few clouds are scattered across the sky, adding depth to the scene. The image is taken from a distance, providing a panoramic view of the town and its surroundings. The perspective allows for a comprehensive view of the town, the water, and the sky, creating a harmonious blend of urban life and natural beauty.



LongCLIP(Ours)

CLIP

The image captures a vibrant scene from a bustling street. The perspective is from a pedestrian's viewpoint on the sidewalk, immersing the viewer in the city's daily life. The street is lined with a variety of buildings, their architecture hinting at the rich history of the city. Among these structures, a church with a tall bell tower stands out, its presence adding a sense of grandeur to the scene. People are seen walking on the sidewalk, going about their day, adding a dynamic element to the otherwise static urban landscape. The colors in the image are predominantly blue, yellow, and green, reflecting the lively atmosphere of the city. The sky above is a clear blue, suggesting a sunny day, which further enhances the overall vibrancy of the scene.

The image captures a serene scene on a river. Dominating the foreground is a wooden boat, its brown hue contrasting with the greenish-blue of the water. The boat is not alone in the frame. In the background, a red-roofed building can be seen, its white walls standing out against the greenery. The building is partially obscured by trees, adding an element of mystery to the scene. The sky above is overcast, casting a soft light over the entire scene. Despite this, there's a sense of tranquility that pervades the image, as if inviting the viewer to take a moment and appreciate the peacefulness of the scene. There's no text or discernible action in the image, just a snapshot of a moment, frozen in time.



LongCLIP(Ours)

CLIP X



Fig. 2: More examples on Image-Text retrieval. The detailed attributes in the long caption to distinguish the correct image is marked in brown.

Text-to-Image Generation

In this photo, a cute dog is relaxing on the grass. He looked very comfortable with his eyes closed, enjoying the warm sun and the gentle breeze. The dog's hair is soft and fluffy, with a warm brown color that makes you want to pet it. On the surrounding grass, a few wildflowers dot it, adding a bit of natural atmosphere.



CLIP



LongCLIP(Ours)

This photograph captures a breathtaking landscape. In the frame, a towering mountain peak stands against the sky, its summit covered in dazzling white snow. Below, a dense forest stretches out towards the horizon, its leaves displaying a riot of colors under the sun's rays. A clear stream trickles through the woods, its melody echoing through the air. In the distance, a few fluffy white clouds float gracefully in the sky, adding a touch of poetry to this breathtaking scene.



In this photograph, an elegant cat lies lazily on the windowsill. Its fur is soft and silky, exuding a charming glow in the sunlight. Outside the window, a lush garden filled with blooming flowers adds a vibrant touch to the scene. A gentle breeze brings the scent of flowers, mingling with the faint aroma of the cat, creating a soothing atmosphere.



CLIP

LongCLIP(Ours)

Fig. 3: More examples on Text-to-Image Generation. The caption marked in brown are the detailed attributes missed by CLIP, but successfully captured by us.

References

- Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12M: Pushing webscale image-text pre-training to recognize long-tail visual concepts. In: CVPR (2021)
- Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793 (2023)
- Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. **123**(1), 32–73 (2017)
- 5. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
- Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., Yan, J.: Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In: ICLR. OpenReview.net (2022)
- Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV. Lecture Notes in Computer Science, vol. 8693, pp. 740–755. Springer (2014)
- Ordonez, V., Kulkarni, G., Berg, T.L.: Im2text: Describing images using 1 million captioned photographs. In: Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F.C.N., Weinberger, K.Q. (eds.) NeruIPS. pp. 1143–1151 (2011)
- Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: improving latent diffusion models for high-resolution image synthesis. CoRR abs/2307.01952 (2023)
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) ICML. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021)
- Wang, J., Zhou, P., Shou, M.Z., Yan, S.: Position-guided text prompt for visionlanguage pre-training. In: CVPR. pp. 23242–23251. IEEE (2023)
- 12. Zeng, Y., Zhang, X., Li, H.: Multi-grained vision language pre-training: Aligning texts with visual concepts. arXiv preprint arXiv:2111.08276 (2021)
- Zeng, Y., Zhang, X., Li, H., Wang, J., Zhang, J., Zhou, W.: X²-vlm: All-in-one pretrained model for vision-language tasks. arXiv preprint arXiv:2211.12402 (2022)