Vision Model Pre-training on Interleaved Image-Text Data via Latent Compression Learning

Chenyu Yang[1,3*], Xizhou Zhu[1,2*], Jinguo Zhu[3,6*], Weijie Su[2,5], Junjie Wang[3,7], Xuan Dong[1,3], Wenhai Wang[2,4], Lewei Lu[3], Bin Li[5], Jie Zhou[1], Yu Qiao[2], Jifeng Dai[1,2✉]

[1]Tsinghua University [2]OpenGVLab, Shanghai AI Laboratory
[3]SenseTime Research [4]The Chinese University of Hong Kong
[5]University of Science and Technology of China [6]Xi'an Jiaotong University
[7]Beijing University of Posts and Telecommunications
{yangcy23,x-dong21}@mails.tsinghua.edu.cn,
{zhuxizhou, jzhou, daijifeng}@tsinghua.edu.cn,
lechatelia@stu.xjtu.edu.cn, jackroos@mail.ustc.edu.cn, jjwang@bupt.edu.cn,
whwang@ie.cuhk.edu.hk, binli@ustc.edu.cn, qiaoyu@pjlab.org.cn, luotto@sensetime.com

Abstract

Recently, vision model pre-training has evolved from relying on manually annotated datasets to leveraging large-scale, web-crawled image-text data. Despite these advances, there is no pre-training method that effectively exploits the interleaved image-text data, which is very prevalent on the Internet. Inspired by the recent success of compression learning in natural language processing, we propose a novel vision model pre-training method called Latent Compression Learning (LCL) for interleaved image-text data. This method performs latent compression learning by maximizing the mutual information between the inputs and outputs of a causal attention model. The training objective can be decomposed into two basic tasks: 1) contrastive learning between visual representation and preceding context, and 2) generating subsequent text based on visual representation. Our experiments demonstrate that our method not only matches the performance of CLIP on paired pre-training datasets (e.g., LAION), but can also leverage interleaved pre-training data (e.g., MMC4) to learn robust visual representations from scratch, showcasing the potential of vision model pre-training with interleaved image-text data. Code is released at https://github.com/OpenGVLab/LCL.

1 Introduction

Over the past decade, ImageNet [34] pre-trained vision models have significantly advanced computer vision, continuously achieving breakthroughs in various vision tasks [7, 24, 10]. The success of ImageNet has inspired further exploration of better methods for pre-training vision models from scratch. Recently, the focus of pre-training has shifted from manually annotated data to large-scale, web-crawled image-text data. A key milestone in this shift is CLIP [55], which utilizes image-text pair data hundreds of times larger than ImageNet, delivering superior performance across various tasks and progressively becoming the mainstream method for vision model pre-training. Building on this trend, there is increasing interest in exploring interleaved image-text data, which is more prevalent on the Internet. Unlike the structured image-text pairs used in CLIP, this interleaved data is free-format and non-paired, larger in scale, and richer in textual information. Fully exploiting these interleaved image-text data is necessary for further improving vision model pre-training at scale.

*Equal contribution. This work is done when Chenyu Yang, Jinguo Zhu, Junjie Wang and Xuan Dong are interns at SenseTime Research.✉Corresponding author: Jifeng Dai <daijifeng@tsinghua.edu.cn>.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-0754

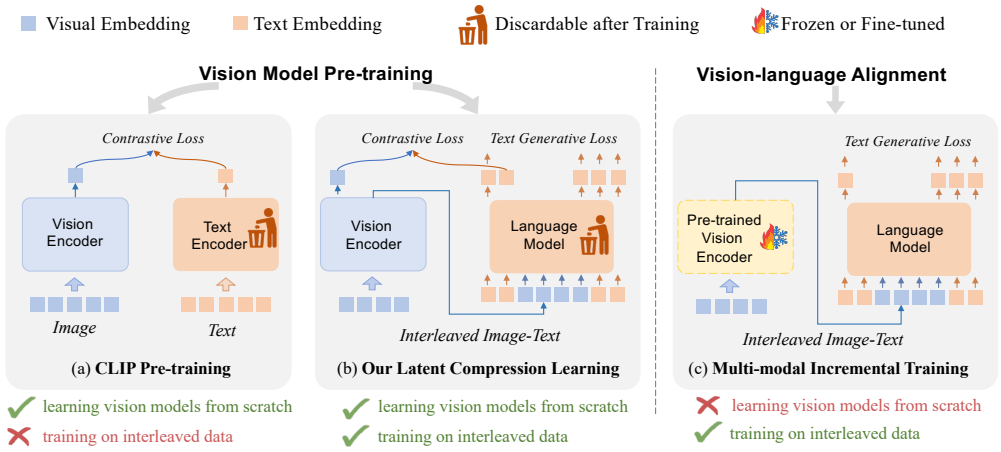


Figure 1: **Comparison of different training frameworks. (a)** Contrastive learning framework from CLIP [55] pre-trains vision encoders from scratch with image-text pairs, but it does not support interleaved data. **(b)** Our proposed LCL pre-training frameworkm can effectively pre-train vision encoders from scratch with interleaved image-text data. In these two frameworks, the text encoder or the language model that provides supervision can be optionally discarded during the transfer stage. **(c)** Multi-modal incremental training process uses interleaved image-text data to align the pre-trained vision encoder and the language model, but it cannot pre-train vision encoders from scratch.

Currently, no pre-training method can effectively utilize interleaved image-text data to pre-train vision models from scratch. In preliminary attempts [2, 29, 44] of using interleaved data, vision models were already pre-trained by CLIP on paired image-text data. Subsequent training on interleaved data primarily serves to align the pre-trained vision models with language models, thereby enhancing the multi-modal capabilities of the entire vision-language network. Therefore, it remains an important and open problem of how to effectively learn robust visual representation from scratch on interleaved image-text data.

A recent study [30] in Natural Language Processing (NLP) suggests that the success of modern language models originates from the compression of training datasets into model parameters. We believe that such compression learning is also applicable to the multi-modal field, except that the data to be compressed expands from structured plain texts to interleaved image-text data, where the images are of raw pixels and unstructured. Such compression learning should be revised to accommodate to the image data. Raw pixels are unstructured and often contain unnecessary and unpredictable details. Such details are irrelevant to the high-level semantic tasks and should be discarded in compression learning. Thus, we argue that compression learning on interleaved text-image data should be applied to latent image representation to better extract semantic abstracts.

In this paper, we propose a novel visual pre-training framework, named Latent Compression Learning. We first theoretically demonstrate that effective latent compression learning can be performed by maximizing the mutual information between outputs and inputs of a causal attention model. When applied to visual pre-training on interleaved image-text data, visual latents are extracted through a visual encoding network (such as ViT [20]), and then fed together with the text into a causal model. The optimization objective can be derived and decomposed into two parts: 1) *contrastive learning* between visual latent representation and their previous context to enhance semantic consistency, and 2) *auto-regressive prediction* to learn the predictability of visual representation for subsequent text. These two training objectives complement each other. For images, the learned latent representation retain information that can be predicted from previous contexts and information needed for predicting subsequent contexts, thus providing effective visual pre-training.

In the experiments, various interleaved and paired pre-training methods are evaluated. The evaluation is conducted through transfer learning on multiple tasks, including image classification, image-text retrieval, image captioning, and visual dialogue. The pre-training datasets include the widely used image-text paired LAION-400M [57] and the image-text interleaved MMC4 [88] and Obelics [36]. In addition, we also re-organize an interleaved version of LAION-Random and a paired version of MMC4-Pair, to facilitate the comparison between interleaved and paired pre-training methods under the same data source. Experiment results show that our LCL pre-training can achieve the same

https://doi.org/10.52202/079017-0754

performance as CLIP on paired pre-training data and can better utilize interleaved pre-training data. Our results also demonstrate the effectiveness of using interleaved image-text data to learn robust visual representation from scratch, and the potential of compression learning for visual pre-training.

2 Related Work

Vision-centric Pre-training Methods. Supervised Pre-training on large-scale annotated datasets [24, 62, 48, 82] has remained the mainstream method for a long time, and has been favored by various vision tasks [9, 24, 79, 7] demonstrating strong performance. Self-Supervised Pre-training has gained significant popularity due to its advantage of utilizing unlabeled data. BEiT [6] follows the methodology of BERT [19] by randomly masking image tokens and reconstructing them as targets. MAE [27] and SimMIM [80] directly use masked pixels as reconstruction targets, making the pre-training process more straightforward and efficient. Weakly-Supervised Pre-training leverages image-hashtag [50, 76, 63] and image-text datasets [71, 61, 8, 57], which rely on noisy text supervision from the internet. For image-hashtag datasets, related works [50, 63] have shown comparatively good performance across various transfer learning settings. In the case of image-text datasets, early efforts [3, 42, 49, 64, 66, 67, 70, 13, 39] focused on learning general visual-linguistic representation. Recently, exemplified by CLIP [55], methods [55, 32] have been developed that involve pre-training through aligned visual-linguistic representation, achieving outstanding results in image classification tasks. Other works like M3I Pre-training [65] propose a unified framework that integrates multiple pre-training strategies with data from various modalities and sources. Currently, weakly supervised pre-training from web-scale text supervision has become the core of multi-modal understanding, but existing methods have not yet to leverage the most widespread interleaved image-text data for training visual representation from scratch.

Interleaved Image-Text Incremental Pre-training. Training using Interleaved Image-Text Data (IITD) has recently garnered significant attention due to the vast amount of such data available online. Recent works[87, 68, 69, 23, 4] such as Flamingo [2] and KOSMOS-1 [29] perform incremental learning on non-public IITD based on previously pre-trained vision and language model parameters as initialization, training text generation models with multi-modal understanding capabilities. With the continuous advancement in the field and the proliferation of public IITD (*e.g.*, MMC4 [88] and OBELICS [36]), numerous models [83, 85, 14] capable of multi-modal understanding have emerged. However, these efforts only perform incremental pre-training on IITD and only analyze the usage of IITD on multi-modal dialogue models. Whether IITD contributes to learning robust visual representation from scratch remains unknown.

Compression Learning in NLP. The perspective [59, 60] that compression is closely connected to intelligence has a long history. A common compression method is arithmetic coding [56, 52], which is a practical approach that uses probability models for optimal data encoding. Some studies [28, 51, 38, 31, 37] argue that compression and intelligence are fundamentally equivalent. With the popularity of large language models, the equivalence of language modeling and compression has once again drawn widespread attention, prompting numerous explorations. Recently, [18] demonstrated through examples that language models serve as universal compressors. [30] posited that language modeling is equivalent to compressing a dataset into model parameters and proposed that compression efficiency is linearly related to model capabilities. Although compression learning has been proven effective in the field of NLP, it is not clear whether this approach can be extended to other fields.

Multi-modal Large Models. Text supervision pre-trained vision models [55, 32, 40] are widely utilized and have exhibited superior performance in tasks ranging from image retrieval and image classification to captioning. Currently, the most popular and closely watched application area is multi-modal dialogue. Multi-modal dialogue models [2, 41, 86, 17, 25, 21, 53] primarily rely on a powerful pre-trained vision encoder [55, 22] and text decoder [16, 74, 75]. The usage of pre-trained vision encoder can generally be divided into two categories: the majority [2, 72], represented by LLaVA [46], employ a strategy where the pre-trained vision encoder is frozen, and only the subsequent adapters and language models are trained. A minority, exemplified by Qwen-VL [5, 53], utilize high-quality image-text dialogue data to continue fine-tuning the vision model. These two training strategies correspond to the two evaluation methods used in this paper to assess the performance of vision models.

https://doi.org/10.52202/079017-0754

3 Method

3.1 Latent Compression Learning

Auto-regressive Language Modeling as Compression Learning. Recent works [18, 30] have shown that auto-regressive language modeling is equivalent to compression learning. Suppose g_ϕ is a language model (LM) with learnable parameters ϕ. Given an input text sequence $x = (\texttt{</s>}, x_1, x_2, \ldots, x_N)$, where </s> is a special token indicating the beginning of text, the model outputs $y = g_\phi(x) = (y_1, y_2, \ldots, y_N)$ predicting the next token based on preceding context, *i.e.*, $\hat{x}_k = y_k = g_\phi(x)_k$. The approximate probability of x estimated by g_ϕ is $q(x) = \prod_{k=1}^{N} q\left(x_k | y_k = g_\phi(x)_k\right)$. The model is optimized with NLL-loss, which equals to minimizing the the cross-entropy between the data distribution p and model distribution q:

$$H(p, q) = \mathbb{E}_{x \sim p}\left[-\sum_{k=1}^{N} \log q\left(x_k | y_k = g_\phi(x)_k\right)\right]. \tag{1}$$

Notice that $H(p, q)$ is actually the optimal expected code length encoding p by q, minimizing $H(p, q)$ just means compressing the data into the model parameters.

Latent Compression for Interleaved Image-Text Data. We believe that the compression principle could apply to multi-modal domain, specifically, to train vision-language models by compressing interleaved image-text data. However, instead of directly dealing with pixel values, we turn to compress high-level image representation for the following reasons: 1) high-level representation can extract useful information from raw pixels while discarding those unpredictable image details. 2) the learned visual representation will align with text semantics, making it possible to perform effective visual pre-training with interleaved image-text data.

Specifically, let $x = (\texttt{</s>}, x_1, x_2, \ldots, x_N)$ be an interleaved image-text sequence. To simplify the expression without loss of generality, we assume that there is only one image in the sequence. Sub-sequence $x_{i:i+M}$ are $M + 1$ image patch tokens of the input image (e.g., non-overlapping patches in ViTs) and others are text tokens. $I = \{i, i + 1, \ldots, i + M\}$ denotes the indices of the image patches, and $T = \{1, \ldots, i - 1, i + M + 1 \ldots, N\}$ denotes the indices of text tokens. As shown in Fig. 2, to construct the sequence of latent representation $z = (\texttt{</s>}, z_1, z_2, \ldots, z_N)$, for image patches, we use a parametric vision encoder f_θ (e.g., ViTs) to map the data sequence $x_{i:i+M}$ into latent variable $z_{i:i+M}$. For text tokens, we directly use one-hot vectors corresponding to their vocabulary ids as the latent codes. Then, the latent representation z are fed into a causal attention model g_ϕ for latent compression by minimizing

$$H(p, q) = -\int p(z) \log q(z) = \mathbb{E}_{x \sim p}\left[-\sum_{k=1}^{N} \int p(z_k | x) \log q\left(z_k | y_k = g_\phi \circ f_\theta(x)_k\right)\right], \tag{2}$$

where the k-th element of the output $y_k = g_\phi \circ f_\theta(x)_k$ predicts the next input latent z_k, f_θ is identity for text tokens for simplicity of annotation. When the compression only applied on text tokens, it degenerates to auto-regressive language modeling on interleaved image-text data used by previous methods (e.g., Kosmos and Flamingo).

However, direct optimizing Eq. (2) for learning informative latent representation is non-trivial, since Eq. (2) suffers from a naturally trivial solution of visual representation collapse, i.e. the image latent representation $z_{i:i+M}$ may be learned to be data-independent. In fact, as showed in Sec. 4.2, we have observed such visual representation collapse, when training from scratch on MMC4 dataset with Eq. (2) applied to text tokens only (i.e., auto-regressive language modeling).

Maximizing Mutual Information for Latent Compression Learning. Optimizing directly for latent compression in Eq. (2) may cause the visual representation collapse. A natural constraint is to maximize the representation entropy to prevent collapse. We find that combining latent compression and maximum entropy constraint is exactly equivalent to maximizing the mutual information between the model inputs and outputs.

Prior work [65] have shown the relationship between cross-entropy and mutual information in other pre-training tasks. Here, we derive this relationship in the latent compression task: maximizing the mutual information between the output y and the input latent z of the causal attention model g_ϕ is

equivalent to compressing z by minimizing $H(p,q)$ in Eq. (2) meanwhile maximizing the entropy of each element z_k in z:

$$\begin{aligned} I(y;z) &= \mathbb{E}_{x\sim p}\left[\sum_{k=1}^{N}\int p(y_k|x)p(z_k|x)\log\frac{p(z_k|y_k)}{p(z_k)}\right] \\ &= \max_q \mathbb{E}_{x\sim p}\left[\sum_{k=1}^{N}\int p(z_k|x)\log q\left(z_k|y_k=g_\phi\circ f_\theta(x)_k\right)\right] - \sum_{k=1}^{N}\int p(z_k)\log p(z_k) \\ &= -\min_q H(p,q) + \sum_{k=1}^{N} H(z_k), \end{aligned} \tag{3}$$

where we use $p(y_k,z_k|x)=p(y_k|x)p(z_k|x)$ in the first step since z_k and y_k can be independently computed given input x, and $p(z_k|y_k)$ is estimated by an approximate parameterized distribution $q(z_k|y_k)$. For the derivation of the formula, please refer to [65].

Therefore, using $I(y;z)$ as the optimization objective can achieve latent compression while avoiding representation collapse of z via the maximum entropy constraint. The compression of z imposes the model to extract useful information and discard unpredictable information of the image. Meanwhile, maximizing $I(y;z)$ requires that each y_k could obtain enough information from previous latent $z_{<k}$ to predict z_k. Each z_k should carry predictable information. These guarantee that the image representation encode rich semantic information aligned with text. We suppose that the above properties learned by the image representation are desired for vision-language pre-training , thus we use Eq. (3) as our pre-training objective. Parameters ϕ and θ are be jointly optimized under this objective. Intuitively, the vision encoder f_θ learns to represent images by high-level abstract, and the causal attention model g_ϕ learns to compress this high-level abstract of the dataset.

3.2 Training Loss

In this sub-section, we demonstrate how Eq. (3) is decomposed into training tasks and losses. Firstly, $I(y;z)$ can be decomposed as a cross-entropy term and an entropy term in the following two symmetric ways (see Appendix B for detailed derivation):

$$I(y;z) = \sum_{k=1}^{N} -\min_{q_1}\mathbb{E}_{x\sim p}\left[H\left(\delta\left[z_k=f_\theta(x)_k\right], q_1\left(z_k|y_k=g_\phi\circ f_\theta(x)_k\right)\right)\right] + H(z_k), \tag{4}$$

$$I(y;z) = \sum_{k=1}^{N} -\min_{q_2}\mathbb{E}_{x\sim p}\left[H\left(\delta\left[y_k=g_\phi\circ f_\theta(x)_k\right], q_2\left(y_k|z_k=f_\theta(x)_k\right)\right)\right] + H(y_k). \tag{5}$$

Since given the input x, latent z_k and y_k are independent and deterministic (*i.e.*, determined by f_θ and g_ϕ), yielding $p(y_k,z_k|x)=\delta\left[z_k=f_\theta(x)_k\right]\cdot\delta\left[y_k=g_\phi\circ f_\theta(x)_k\right]$. $\delta[\cdot]$ is delta distribution. In Eq. (4), $p(z_k|y_k)$ is estimated by a parameterized distribution $q_1(z_k|y_k)$, which approximates the distribution of z_k given the model's prediction y_k. Similarly, in Eq. (5), $p(y_k|z_k)$ is estimated by $q_2(y_k|z_k)$, the predicted distribution of y_k given z_k. Therefore, maximizing mutual information can be decomposed as follows: 1) The causal attention model g_ϕ learns to predict the next latent z_k from the output y_k. 2) The learnable latent representation z_k learns to predict y_k, which is the representation of its previous context. 3) The maximum entropy regularization avoids the collapse of z_k and y_k.

In the following, we show that the cross-entropy terms in $I(y;z)$ can be achieved by two common training tasks and loss functions, while the entropy constraints are implicitly satisfied.

Contrastive Learning between Image Representation and Preceding Context. For image latent z_k and the corresponding y_k representing the semantics of its preceding context, the objective defines a bidirectional prediction. We choose q as Boltzmann distribution, *i.e.*, $q\left(z_k|y_k\right)\propto\exp(z_k^\top W_1^\top W_2 y_k/\tau)$ and $q\left(y_k|z_k\right)\propto\exp(y_k^\top W_2^\top W_1 z_k/\tau)$, where τ is the temperature, W_1 and W_2 are learnable linear projections. Consequently, the objective becomes the contrastive loss in two directions between z_k and y_k, when setting $z_{k'}$ and $y_{k'}$ from other images as negative samples:

$$L_{con} = -\sum_{k\in I}\log\frac{\exp(y_k^\top W_2^\top W_1 z_k/\tau)}{\sum_{k'}\exp(y_k^\top W_2^\top W_1 z_{k'}/\tau)} - \sum_{k\in I}\log\frac{\exp(z_k^\top W_1^\top W_2 y_k/\tau)}{\sum_{k'}\exp(z_k^\top W_1^\top W_2 y_{k'}/\tau)} \tag{6}$$

https://doi.org/10.52202/079017-0754

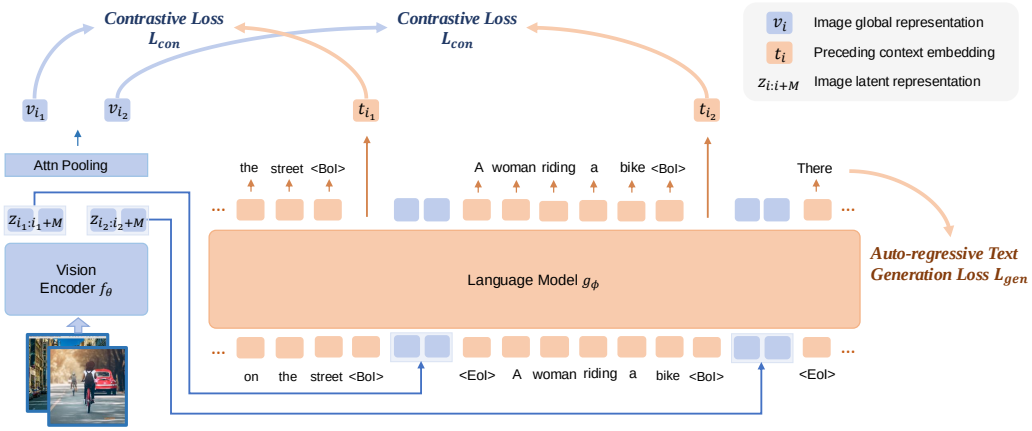


Figure 2: **Overview of our proposed Latent Compression Learning for vision model pre-training.** Image latent representation is extracted via a vision encoder and subsequently input into a language model alongside textual embedding. Two complementary losses are utilized to learn robust visual representation from scratch on interleaved image-text data: a contrastive loss ensures consistency between the visual latent representation and its preceding context, while an auto-regressive loss enhances the predictability of visual representation for subsequent text.

Meanwhile, the contrastive loss also prevents z_k and y_k from being trivial representation by pulling them away from negative samples, implicitly appending the entropy regularization.

Auto-regressive Text Generation. For a text token, its latent code z_k is a one-hot vector and is not learnable, so the objective only imposes y_k to predict z_k as in Eq. (4). We choose $q(z_k|y_k)$ as softmax over the output logits on the text vocabulary, i.e, $q(z_k|y_k) = z_k^\top \text{softmax}(V y_k)$, where V is the projection head of the language model. The objective corresponding to the text tokens is simply the objective of standard next token prediction with cross-entropy loss:

$$L_{gen} = -\sum_{k \in T} \log z_k^\top \text{softmax}(V y_k) \tag{7}$$

The total training loss is defined as $L = \lambda L_{con} + L_{gen}$, where λ is balancing weight.

Relation to Previous Pre-training Tasks. In our proposed pre-training framework, the contrastive task and generation task align the image representation with preceding context and subsequent context, respectively. Hence, combining these two pre-training tasks can fully leverage the semantic information contained in interleaved image-text data to supervise the learning of image representation. For previous pre-training tasks, 1) *Contrastive Language-Image Pre-training (CLIP)* [55] has a similar objective of maximizing the mutual information between corresponding image and text [65], but it can only be applied to paired image-text data. 2) *Contrastive Captioner (CoCa)* [81] combines a captioning (text generation) loss with the CLIP loss, but it cannot be applied to interleaved data, either. CLIP loss requires image-text pairs, while the captioner can only be conditioned on a single image can rather than flexible interleaved image-text contents. 3) *Auto-regressive Text Generation* task only leverages the semantic information in subsequent context to supervise the learning of image representation, while the preceding context is missing, and representation collapse cannot be avoided. Moreover, interleaved image-text data usually has information redundancy, *i.e.*, the image and its corresponding text may contain similar information. So models may rely on information from the text rather than the image for prediction. This is particularly true when the vision encoder is trained from scratch, resulting that the image representation are never focused and optimized. Our experiments in Sec. 4.2 confirms this analyze.

3.3 Architecture

The overview of model architecture when adopting our LCL is shown in Fig. 2. The interleaved image-text input sequence may contain multiple images. We adopt a Vision Transformer (ViT) [20] as the vision encoder, which encodes each image into a sequence of visual embeddings as its latent representation. The visual embeddings of each image are inserted at corresponding positions in the

https://doi.org/10.52202/079017-0754

interleaved sequence, and we introduce special tokens <BoI> and <EoI> to indicate the beginning and the ending positions, respectively. The combined visual and text embeddings are fed into a causal language model. The text generation loss is the standard cross-entropy loss over the output logits of the language model defined in Eq. (7). In the contrastive learning task, calculating the loss for each image token is extremely computationally expensive. To alleviate this problem, we consider utilizing one global representation per image instead of all latent representation for contrastive learning. Specifically, the global representation v_i of each image is extracted from its latent representation $z_{i:i+M}$ through an attention pooling, which is a multi-head attention layer with a learnable query token. The output of the causal transformer model at <BoI> token, just before the image, is processed by a LayerNorm layer and a linear projection into t_i. The contrastive loss in Eq. (6) is calculated between v_i and t_i.

4 Experiment

4.1 Experiment Settings

Pre-train Data. The datasets utilized in our pre-training encompass the image-text pair dataset LAION-400M [57], as well as the image-text interleaved datasets MMC4 [88] and OBELICS [36]. We also re-organize two datasets for fair comparison with CLIP. 1) LAION-Random as a interleaved dataset. Images from LAION-400M are randomly placed before or after their paired caption to form image-text sequences. 2) MMC4-Pair as a paired dataset. For images in MMC4, we select matching text from the bipartite graph matching results derived from CLIP similarity to generate pseudo image-text paired data.

Implementation Details. We adopt the same image transform in OpenCLIP [15] for pre-training and set the image size as 224×224 for all experiments. We employ ViT [20] as our vision encoder. ViT-L/14 is used for the main results, and ViT-B/16 is used for ablation studies. The language model follows the same architecture as OPT-125M [84] but is randomly initialized. By default, the contrastive loss balancing weight is set at $\lambda = 0.1$. AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$ and a weight decay of 0.1 are used. We employ a cosine learning rate schedule with linear warmup and set the peak learning rate at 5e-4 for the LAION data and 3e-4 for others. The model for the main results is trained for 200k iterations with 64k images per batch on average. In the ablation study, the models are trained for 250k iterations with 8k images per batch.

Vision Encoder Evaluation. CLIP pre-trained models typically uses zero-shot retrieval performance for evaluation. However such zero-shot inference on retrieval tasks [55] is not suitable for other generative pre-training methods. To address this discrepancy, we choose to evaluate pre-trained vision models by transfer learning on various downstream tasks under two configurations: "frozen transfer" and "full transfer". In the frozen transfer setting, only the parameters outside of the vision model are trained. In the full transfer setting, all parameters are trained.

The evaluation is conducted on image classification, image-text retrieval, image caption, and multi-modal dialogue. Please see Appendix A for more data, implementation, and evaluation details.

4.2 Comparison with Vision Pre-training Methods

In this section, we show the superiority of our LCL when using interleaved image-text data for vision pre-training. However, existing vision pre-training methods only support paired data, so we choose MMC4 dataset with a paired set (MMC4-Pair) available for those methods. In addition, to confirm that it is is not trivial to use interleaved data for vision pre-training, we also consider existing methods to use interleaved data. Those methods are originally proposed as MLLM training methods, but here we test them for vision pre-training. Besides, on paired image-text data, we also compare our LCL with existing vision pre-training methods and show they are comparable. See Appendix A.3 and A.2 for detailed pre-training and evaluation settings.

We aggregate all the pre-training methods divide them into the following tasks: (1) Image-text contrastive (Con.) [55, 15, 32], (2) Image-text contrastive + image captioning (Con. + Cap.) [81], (3) Image-text contrastive + image captioning + image-text matching (Con. + Cap. + Mat.) [40, 41, 14], (4) Auto-regressive text generation (Gen.) [2, 29, 44], (5) Auto-regressive text generation + image regression (Gen. + Reg.) [69, 68, 72], (6) Mask data modeling (Mask.) [78]. We select one

https://doi.org/10.52202/079017-0754

Table 1: **Frozen transfer evaluations of vision models pre-trained on the MMC4 dataset.** Vision models are pre-trained from scratch for all methods. "IN-1k" denotes image classification on ImageNet [33]. "ret." and "cap." denote image-text retrieval and image captioning, respectively. * The method names refer to implementing those methods with our experiment setting but not their trained checkpoints. For pre-trainig tasks, **Con.** image-text contrastive; **Cap.** image captioning; **Mat.** image-text matching; **Mask.** mask data modeling; **Gen.** auto-regressive text generatiton; **Reg.** image (feature) regression. The names in parentheses refer to the pre-training task but not their trained checkpoints. † Note that CoCa and BLIP2 need to pass each sample through the language model 2 and 3 times, respectively, to perform multi-task learning.

* Pre-training method (task)	Pre-training data	IN-1k	COCO ret.		Flickr30k ret.		COCO cap.		NoCaps cap.	
		acc-1	TR@1	IR@1	TR@1	IR@1	B@4	C	B@4	C
CLIP (Con.)	MMC4-Pair	74.8	46.4	32.5	**76.2**	60.0	23.9	82.9	29.6	78.4
†CoCa (Con. + Cap.)	MMC4-Pair	**75.4**	**48.6**	**34.3**	**76.5**	**61.9**	23.7	84.8	30.0	80.5
†BLIP2 (Con. + Cap. + Mat.)	MMC4-Pair	74.5	46.5	31.3	74.9	57.8	23.7	82.9	29.4	78.1
BEiT3 (Mask.)	MMC4	73.3	45.1	30.6	73.2	57.1	23.3	81.4	29.5	76.7
Flamingo (Gen.)	MMC4	24.0	10.6	5.6	17.7	10.8	8.7	18.1	15.0	18.5
Emu (Gen. + Reg.)	MMC4	5.7	2.3	1.4	4.8	2.7	0.3	4.4	0.6	4.6
LCL (Ours)	MMC4	**75.2**	**48.5**	**34.5**	**76.3**	60.4	**24.4**	**87.5**	**31.0**	**82.5**

Table 2: **Frozen transfer evaluations of vision models pre-trained on LAION dataset.** Vision models are pre-trained from scratch for all methods. * The method names refer to implementing those methods with our experiment setting but not their trained checkpoints. † Note that CoCa and BLIP2 need to pass each sample through the language model 2 and 3 times, respectively, to perform multi-task learning.

* Pre-training method (task)	Pre-training data	IN-1k	COCO ret.		Flickr30k ret.		COCO cap.		NoCaps cap.	
		acc-1	TR@1	IR@1	TR@1	IR@1	B@4	C	B@4	C
CLIP (Con.)	LAION-400M	**75.0**	47.2	34.2	76.5	59.8	24.1	84.1	30.0	78.8
†CoCa (Con. + Cap.)	LAION-400M	**75.2**	**48.6**	**34.8**	**76.9**	**61.4**	**24.6**	**88.2**	30.4	82.9
†BLIP2 (Con. + Cap. + Mat.)	LAION-400M	74.0	47.4	31.5	75.9	57.9	23.6	85.0	30.0	77.8
LCL (Ours)	LAION-Random	**75.1**	**48.3**	34.3	**76.8**	59.6	**24.4**	**88.1**	**31.3**	**84.2**

representative method from each task. For fair comparison, we implement these methods on the same model and training data with available open-source codes or our reproduction.

Pre-training on Interleaved Image-Text Data. We conduct experiments on the MMC4 dataset [88] to demonstrate the effectiveness of our LCL on interleaved image-text data. For pre-training methods that only support paired data, MMC4-Pair is used.

The results are shown in Tab. 1. Our LCL pre-training method significantly outperforms all other methods in the caption tasks, indicating that we can effectively utilize the rich text context information in MMC4 interleaved data. On the other hand, our method is on par with the best paired pre-training methods on classification and retrieval tasks. Since the paired pre-training methods are directly optimized for retrieval, our comparable performance shows that the visual features are learned to be highly distinguishable. It is worth mentioning that for more general interleaved data, where no paired versions exist, these paired pre-training methods cannot be applied.

In addition, we observed that the two methods using auto-regressive text generation do not achieve good performance and feature collapse occurs. However, their text prediction training loss is actually close to ours. This suggests that these methods tend to rely on redundant text information rather than image information for subsequent text prediction. As discussed in Sec. 3.1, our approach can avoid such collapse.

Pre-training on Paired Image-Text Data. We conduct experiments on the LAION-400M dataset to show that LCL pre-training also performs well on paired image-text data without specific modification. Tab 2 shows that our method is comparable to paired pre-training methods on various tasks, indicating that all information in paired data is fully exploited.

https://doi.org/10.52202/079017-0754

Table 3: Transfer evaluation results of pre-trained ViT/L-14 on classification, retrieval and captioning tasks.

Model	Pre-training data	Pre-training epoch	IN-1k	COCO ret.		Flickr30k ret.		COCO cap.		NoCaps cap.	
			acc-1	TR@1	IR@1	TR@1	IR@1	B@4	C	B@4	C
frozen transfer											
OpenAI CLIP	WIT-400M	32	83.7	61.7	48.2	89.0	75.8	32.1	116.0	35.5	108.9
OpenCLIP	LAION-400M	32	**82.1**	59.5	**46.0**	86.9	74.2	31.0	111.5	34.8	106.0
LCL (Ours)	LAION-400M	32	**82.2**	59.6	**46.2**	86.7	74.0	31.4	112.3	**35.0**	106.7
LCL (Ours)	LAION-400M + MMC4	16	**82.0**	**60.0**	**46.0**	**87.6**	**74.6**	**32.0**	**113.7**	**35.1**	**107.1**
full transfer											
OpenAI CLIP	WIT-400M	32	87.4	62.1	49.6	90.3	77.9	39.5	132.7	41.4	116.9
OpenCLIP	LAION-400M	32	**86.2**	61.7	**47.5**	87.7	**76.3**	38.6	128.9	39.9	112.7
LCL (Ours)	LAION-400M	32	**86.1**	**61.9**	**47.5**	87.6	**76.1**	39.1	129.7	40.2	113.5
LCL (Ours)	LAION-400M + MMC4	16	**86.1**	**62.2**	**47.6**	**88.1**	**76.2**	**39.5**	**130.9**	**40.6**	**113.8**

Table 4: Transfer evaluation results of pre-trained ViT/L-14 on multi-modal benchmarks. The transfer adopts the downstream model and training pipeline of LLaVA-1.5 [46].

Model	Pre-training data	Pre-training epoch	VQAv2	GQA	VisWiz	SQA	POPE	MME	MMB	$SEED_I$
frozen transfer										
OpenAI CLIP	WIT-400M	32	77.1	61.7	44.4	71.1	84.6	1486.9	65.1	64.6
OpenCLIP	LAION-400M	32	68.7	57.0	39.5	69.0	**81.8**	1266.2	54.2	55.8
LCL (Ours)	LAION-400M + MMC4	16	**70.7**	**57.4**	**41.4**	**69.7**	**81.7**	**1291.7**	**55.3**	**56.5**
full transfer										
OpenAI CLIP	WIT-400M	32	79.0	62.8	46.8	71.8	85.7	1576.8	68.9	67.9
OpenCLIP	LAION-400M	32	71.5	**58.6**	42.2	70.4	**82.5**	1345.4	58.5	59.5
LCL (Ours)	LAION-400M + MMC4	16	**73.4**	**58.8**	**44.2**	**71.0**	**82.5**	**1382.3**	**59.5**	**60.3**

4.3 Comparison with Pre-trained Checkpoints

To further confirm the effectiveness of our proposed Latent Compression Learning (LCL), we compare our pre-trained model with existing checkpoints of pre-trained vision encoders. We use LCL to pre-train a ViT-L/14 with mixed data from the LAION-400M and MMC4. We compare it to the ViT-L/14 pre-trained by OpenCLIP [15] using the public LAION-400M dataset, and the ViT-L/14 pre-trained by OpenAI CLIP [55] with private data is listed as reference. The total number of images seen during pre-training is 13B for all models. We evaluate the pre-trained vision encoders by transferring to downstream tasks, *i.e.*, integrate the vision encoders into downstream task models and compare the fine-tuning results. More training details are in Appendix A.1 and evaluation details are in Appendix A.2.

Tab. 3 and Tab. 4 show the results of transfer evaluations. When both use LAION-400M as pre-training data, as with the previous experimental conclusions, LCL has similar performance to CLIP. When combined with MMC4, our method achieves better performance, especially on caption and multi-modal dialogue tasks.

There exist approaches achieving better results on some benchmarks. However, they either use larger vision encoders, more training data, or private data. We list those results as reference in Appendix C.

4.4 Ablation Study

Latent Compression Learning on Different Datasets. We apply LCL pre-training to more datasets to confirm its generalizability. As shown in Tab. 5, our method also achieves reasonable performance on interleaved dataset OBELICS. It is worth noting that the models trained on MMC4 and OBELICS have achieved similar performance to that on LAION, indicating that it is completely feasible to pre-train visual models only from interleaved data. Furthermore, using both LAION and MMC4 data during pre-training improves performance, suggesting that further improvements can be obtained by incorporating more image-text data. In this case, supporting interleaved data is a key advantage of our approach, enabling the use of more diverse image-text data for pre-training.

https://doi.org/10.52202/079017-0754

Table 5: **Frozen transfer evaluations of LCL pre-training on different datasets.** We ensured that all entries had seen the same number of images during pre-training to ensure fairness.

Pre-training method	Pre-training data	IN-1k	COCO ret.		Flickr30k ret.		COCO cap.		NoCaps cap.	
		acc-1	TR@1	IR@1	TR@1	IR@1	B@4	C	B@4	C
LCL (Ours)	LAION	75.1	48.3	34.3	76.8	59.6	24.4	88.1	31.3	84.2
	MMC4	75.2	48.5	34.5	76.3	60.4	24.4	87.5	31.0	82.5
	Obelics	73.9	47.0	33.2	75.1	58.7	23.3	84.8	29.9	77.6
	LAION + MMC4	**75.8**	**49.4**	**35.3**	**77.7**	**61.1**	**25.1**	**88.9**	**31.4**	**84.9**

Table 6: **Ablations of the training loss and loss balancing weight in LCL .** Models are evaluated under frozen transfer setting.

(a) Training loss ablation.

Training loss	COCO ret.		COCO cap.	
	TR@1	IR@1	B@4	CIDEr
Con. only	46.4	32.7	23.8	82.7
Gen. only	10.3	5.4	8.5	17.8
LCL	**48.5**	**34.5**	**24.4**	**87.5**

(b) Loss balancing weight ablation.

Con. weight λ	COCO Ret.		COCO Cap.	
	TR@1	IR@1	B@4	CIDEr
0.05	48.3	33.7	24.3	87.0
0.1	**48.5**	**34.5**	**24.4**	**87.5**
0.2	47.6	33.1	24.2	86.3
0.5	37.1	23.2	20.4	66.7

Loss Balance in Latent Compression Learning. Table 6a ablates the contrastive loss and generation loss used in LCL . Consistent with the previous analyses, LCL can achieve the best performance. Table 6b studies the appropriate loss balancing weights (multiplied by the contrastive loss). It turns out that $\lambda = 0.1$ will produce the best results. The performance drops significantly for larger λ values, indicating that the optimization directions of the two losses are not completely consistent.

5 Conclusion

No existing work has explored vision model pre-training with interleaved image-text data. To this end, we propose Latent Compression Learning (LCL) framework that compresses interleaved image-text latent for vision pre-training. We theoretically show that latent compression is equivalent to maximizing the mutual information between the input and output of a causal model and further decompose this objective into two basic training tasks. Experiments demonstrate that our method is comparable to CLIP on paired pre-training datasets, and it effectively learns robust visual representations utilizing interleaved image-text data. Our work showcases the effectiveness of using interleaved image-text data to learn robust visual representation from scratch, and confirms the potential of compression learning for visual pre-training.

Limitations. Our experiments are constrained to a limited size of dataset and vision encoder, and the scaling property of our proposed method remains unexplored.

Acknowledgments.

This work is supported by the National Key R&D Program of China (NO. 2022ZD0161300), by the National Natural Science Foundation of China (62376134).

References

[1] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019.

[2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[3] C. Alberti, J. Ling, M. Collins, and D. Reitter. Fusion of detected objects in text for visual question answering. *arXiv preprint arXiv:1908.05054*, 2019.

[4] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Gadre, S. Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

[5] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. 2023.

[6] H. Bao, L. Dong, S. Piao, and F. Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[7] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[8] S. Changpinyo, P. Sharma, N. Ding, and R. 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.

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[11] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[12] X. Chen, M. Ding, X. Wang, Y. Xin, S. Mo, Y. Wang, S. Han, P. Luo, G. Zeng, and J. Wang. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, 132(1): 208–223, 2024.

[13] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[14] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, Z. Muyan, Q. Zhang, X. Zhu, L. Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

[15] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

[16] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[17] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

[18] G. Delétang, A. Ruoss, P.-A. Duquenne, E. Catt, T. Genewein, C. Mattern, J. Grau-Moya, L. K. Wenliang, M. Aitchison, L. Orseau, et al. Language modeling is compression. *arXiv preprint arXiv:2309.10668*, 2023.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

 https://doi.org/10.52202/079017-0754

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[21] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[22] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023.

[23] Y. Ge, S. Zhao, Z. Zeng, Y. Ge, C. Li, X. Wang, and Y. Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023.

[24] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[25] T. Gong, C. Lyu, S. Zhang, Y. Wang, M. Zheng, Q. Zhao, K. Liu, W. Zhang, P. Luo, and K. Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023.

[26] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[27] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[28] J. Hernández-Orallo and N. Minaya-Collado. A formal definition of intelligence based on an intensional variant of algorithmic complexity. In *Proceedings of International Symposium of Engineering of Intelligent Systems (EIS98)*, pages 146–163, 1998.

[29] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, B. Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[30] Y. Huang, J. Zhang, Z. Shan, and J. He. Compression represents intelligence linearly. *arXiv preprint arXiv:2404.09937*, 2024.

[31] M. Hutter. The human knowledge compression prize. *URL http://prize. hutter1. net*, 2006.

[32] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[35] G. Larsson, M. Maire, and G. Shakhnarovich. Fractalnet: Ultra-deep neural networks without residuals. *arXiv preprint arXiv:1605.07648*, 2016.

[36] H. Laurençon, L. Saulnier, L. Tronchon, S. Bekman, A. Singh, A. Lozhkov, T. Wang, S. Karamcheti, A. Rush, D. Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024.

[37] S. Legg and M. Hutter. Universal intelligence: A definition of machine intelligence. *Minds and machines*, 17:391–444, 2007.

[38] S. Legg, M. Hutter, et al. A universal measure of intelligence for artificial agents. In *International Joint Conference on Artificial Intelligence*, volume 19, page 1509. LAWRENCE ERLBAUM ASSOCIATES LTD, 2005.

[39] G. Li, N. Duan, Y. Fang, M. Gong, and D. Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11336–11344, 2020.

[40] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[41] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[42] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[43] Y. Li, H. Fan, R. Hu, C. Feichtenhofer, and K. He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23390–23400, 2023.

[44] J. Lin, H. Yin, W. Ping, Y. Lu, P. Molchanov, A. Tao, H. Mao, J. Kautz, M. Shoeybi, and S. Han. Vila: On pre-training for visual language models. *arXiv preprint arXiv:2312.07533*, 2023.

[45] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[46] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[47] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[48] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[49] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[50] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.

[51] M. V. Mahoney. Text compression as a test for artificial intelligence. *AAAI/IAAI*, 970, 1999.

[52] R. C. Pasco. *Source coding algorithms for fast data compression*. PhD thesis, Stanford University CA, 1976.

[53] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, and F. Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[54] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[55] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[56] J. J. Rissanen. Generalized kraft inequality and arithmetic coding. *IBM Journal of research and development*, 20(3):198–203, 1976.

[57] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[58] C. Schuhmann, A. Köpf, R. Vencu, T. Coombes, and R. Beaumont. Laion coco: 600m synthetic captions from laion2b-en. *URL https://laion. ai/blog/laion-coco*, 2022.

[59] C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423, 1948.

[60] C. E. Shannon. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64, 1951.

 https://doi.org/10.52202/079017-0754

[61] P. Sharma, N. Ding, S. Goodman, and R. 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.

[62] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[63] M. Singh, L. Gustafson, A. Adcock, V. de Freitas Reis, B. Gedik, R. P. Kosaraju, D. Mahajan, R. Girshick, P. Dollár, and L. Van Der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 804–814, 2022.

[64] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[65] W. Su, X. Zhu, C. Tao, L. Lu, B. Li, G. Huang, Y. Qiao, X. Wang, J. Zhou, and J. Dai. Towards all-in-one pre-training via maximizing multi-modal mutual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15888–15899, 2023.

[66] C. Sun, F. Baradel, K. Murphy, and C. Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019.

[67] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019.

[68] Q. Sun, Y. Cui, X. Zhang, F. Zhang, Q. Yu, Z. Luo, Y. Wang, Y. Rao, J. Liu, T. Huang, et al. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*, 2023.

[69] Q. Sun, Q. Yu, Y. Cui, F. Zhang, X. Zhang, Y. Wang, H. Gao, J. Liu, T. Huang, and X. Wang. Emu: Generative pretraining in multimodality. In *The Twelfth International Conference on Learning Representations*, 2023.

[70] H. Tan and M. Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[71] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[72] C. Tian, X. Zhu, Y. Xiong, W. Wang, Z. Chen, W. Wang, Y. Chen, L. Lu, T. Lu, J. Zhou, et al. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. *arXiv preprint arXiv:2401.10208*, 2024.

[73] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 32–42, 2021.

[74] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[75] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[76] A. Veit, M. Nickel, S. Belongie, and L. Van Der Maaten. Separating self-expression and visual content in hashtag supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5919–5927, 2018.

[77] T. F. Y. Vicente, L. Hou, C.-P. Yu, M. Hoai, and D. Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 816–832. Springer, 2016.

[78] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

[79] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018.

[80] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022.

[81] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[82] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022.

[83] P. Zhang, X. D. B. Wang, Y. Cao, C. Xu, L. Ouyang, Z. Zhao, S. Ding, S. Zhang, H. Duan, H. Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023.

[84] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[85] H. Zhao, Z. Cai, S. Si, X. Ma, K. An, L. Chen, Z. Liu, S. Wang, W. Han, and B. Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*, 2023.

[86] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[87] J. Zhu, X. Ding, Y. Ge, Y. Ge, S. Zhao, H. Zhao, X. Wang, and Y. Shan. Vl-gpt: A generative pre-trained transformer for vision and language understanding and generation. *arXiv preprint arXiv:2312.09251*, 2023.

[88] W. Zhu, J. Hessel, A. Awadalla, S. Y. Gadre, J. Dodge, A. Fang, Y. Yu, L. Schmidt, W. Y. Wang, and Y. Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36, 2024.

 https://doi.org/10.52202/079017-0754

A Experimental Details

A.1 Pre-training

Data. For the data in MMC4, we select images with a CLIP similarity to the matching text of 0.24 or higher. From documents containing at least one such image, we randomly choose up to 6 images to form an interleaved image-text sequence, utilizing all text from that document. If the sequence length exceeds 2048 tokens, the surplus is truncated while ensuring the integrity of both images and individual text segments, and then padded to the designated length. For OBELICS, we similarly restrict the number of images per document to between 1 and 6. We then sequentially extract 2048 tokens from the concatenated documents. If image tokens are truncated, the entire image is moved to the next sample sequence.

To construct interleaved image-text samples from the MMC4 dataset, we randomly place images either before or after their corresponding sentences, adhering to a 50% probability, thus generating a document-wise interleaved sequence of images and text. For the OBELICS corpus, individual documents are concatenated, and a sliding window strategy is employed to select each image-text sequence, maintaining a total length of 2048 tokens.

Hyper-parameters. Our pre-training configuration is shown in Tab. 7. The AdamW optimizer was employed for model training with the learning rate set to 3e-4 and the weight decay set to 0.1. Mixed numerical precision training with bfloat16 is also employed to stabilize the optimization process. Furthermore, we set a drop-path [35] rate linearly increasing to 0.2, and use layer-scale [73] for stable training.

Table 7: **Hyper-parameters in pre-training.**

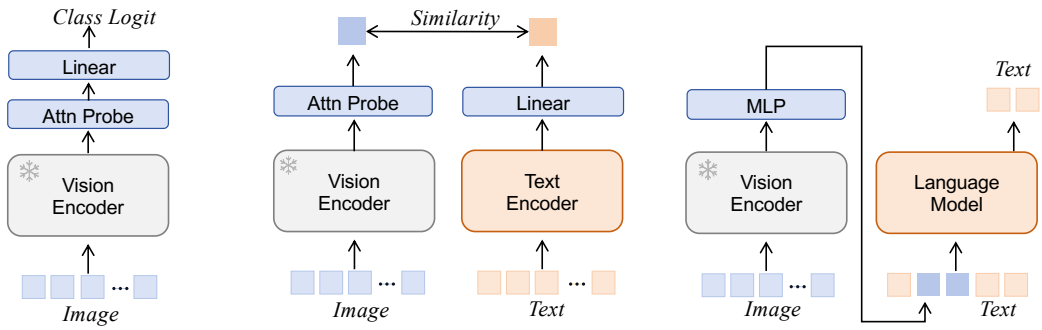
optimizer	AdamW
learning rate	3e-4
weight decay	0.1
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.95$
lr schedule	cosine decay
warmup	2k
numerical precision	bfloat16
train steps	20k
batch size (in images)	64k
drop path	0.2

A.2 Evaluation

Transfer Tasks. We conduct our performance evaluation of pre-trained on image classification, image-text retrieval, text generation tasks with multimodal inputs (*i.e.*, image captioning and multi-modal dialogue). Their model architecture in transfer learning are illustrated in Fig. 3.

For closed-set image classification, a lightweight classifier with a randomly initialized attention pooling layer, followed by a layer normalization layer and a linear layer, is appended to the top of the pre-trained vision model. In the "frozen transfer" scenario, only the parameters of the added classifier are trainable, similar to the linear probing strategy. Conversely, in the "full transfer" approach, all parameters, including those of the pre-trained vision encoder, are adjustable.

Regarding the image-text retrieval task, we discard the pre-trained text encoder and apply contrastive learning to the pretrained vision encoder with a newly introduced text encoder. Images are processed through the vision encoder and a randomly initialized attention pooling layer to generate a global image embedding. Textual captions are processed through the text encoder, utilizing the feature of the final token as the text embedding representing the input caption. An additional linear layer facilitates the dimensional alignment between the image and text embeddings, enabling their use in contrastive learning. During "frozen transfer", the attention pooling layer on the vision transformer and the text encoder are trainable; similarly, in "full transfer", all parameters, including the vision encoder, can be optimized.



(a) Image Classification (b) Vision-Language Retrieval (c) Text Generation

Figure 3: **Illustration of "frozen transfer" evaluation**. The vision encoder is frozen during transfer tuning. (a) Image classification: an attention probe and a linear classifier are built upon the vision encoder. (b) Image-text retrieval: an attention probe is used to extract global visual feature, which is trained to align with the text feature from the text encoder. (c) Text generation: an MLP is utilized to align the visual feature with the text embedding space, and the multi-modal embedding is fed into the language model for auto-regressive text generation.

For text generation tasks with multimodal inputs, we adopt image captioning and multi-modal dialogue benchmarks and employ prevalent architectures like those in [46]. Specifically, a pretrained LLM for text generation is integrated on top of the pre-trained vision model, incorporating an MLP layer to adjust the dimensions of the visual embeddings. During the "frozen transfer" evaluation, the parameters of the vision model remain fixed, while in the 'unfreezing' phase, these parameters are permitted to undergo training.

Implementation Details. Vision encoder evaluation with "frozen transfer" and "full transfer" configurations includes fine-tuning training and benchmark evaluation. Implementation details of each transfer task are list below, and the hyper-parameters involved are listed in Tab. 8.

- *Image classification.* Model is trained on the ImageNet-1K [33] *train* split and evaluated on *val* split. We follow the attention probe setting introduced by [12] for "frozen transfer", and the full fine-tune setting in [43] for "full transfer".
- *Image-text retrieval.* Model is trained on a combination dataset comprised of CC12M [61], CC3M [61], and SBU [77], and is tested on the MSCOCO [11] *karpathy-test* split and Flickr30k [54] *test* split. The model is trained with Adamw optimizer for 5000 iterations. The learning rate is set at 1e-3 and 1e-5 for parameters without initialization and with initialization, respectively.
- *Image captioning.* Model is trained on a subset of the LAION-COCO [58] dataset, which includes 10 million samples, and evaluation is performed on the MSCOCO [11] *karpathy-test* split and NoCaps [1] *val* split. Here, the model is trained for 20,000 iterations with a learning rate of 1e-4. Additionally, a droppath technique is employed with a ratio of 0.2 in the vision model to mitigate overfitting.
- *Multi-modal dialogue.* We follow a two-stage training process similar to that used in LLAVA-1.5 [45]. Initially, paired data with 558K samples is used to train an MLP projector to align the Vision Transformer (ViT) with the pretrained LLM. Subsequently, the model undergoes instruction tuning on multimodal dialogue datasets with 665k samples. Both the alignment training and instruction tuning phases are conducted over a single epoch, with learning rates set at 1e-3 and 2e-5, respectively. Evaluations are then performed on multimodal dialogue benchmark, *e.g.*, MMBench [47] and VQAv2 [26].

A.3 Ablation Experiments

The effectiveness of our LCL are validated by conducting ablation experiments mainly on two corpora: LAION and MMC4. The experimental hyper-parameters involved are shown in Tab. 9. We found that the optimal learning rate for the LAION dataset is 5e-4, while for the MMC4 dataset, a slightly lower

https://doi.org/10.52202/079017-0754

Table 8: **Hyper-parameters in transfer evaluation.**

Hyper-parameter	Classification	Retrieval	Image Captioning	Multi-modal dialogue	
				stage1	stage2
train dataset	IN-1K train	CC12M, CC3M, SBU	LAION-COCO	LCS-558k	LLaVA-SFT
test dataset	IN-1K val	COCO,Flickr30k	COCO,Nocaps	MMbench, VQAv2, GQA...	
optimizer	AdamW				
learning rate	1e-4	1e-3	1e-4	1e-3	2e-5
weight decay	1e-4				
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.95$				
learning rate schedule	cosine decay				
warmup	1500	500	1000	30	156
train steps	14k (90ep)	5k	20k	1091	5198
batch size (in images)	8192	16k	512	512	128

rate of 3e-4 proves most effective. We speculate that this is because MMC4 corpus contains relatively higher noise. Most of the original settings in the large-scale pre-training are retained in the ablation, with the exception of reducing the batch size by a factor of 8 to decrease the computational overhead.

Table 9: **Hyper-parameters in ablation study.**

Hyper-parameter	LAION	MMC4
optimizer	AdamW	
learning rate	5e-4	3e-4
weight decay	0.1	
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.95$	
learning rate schedule	cosine decay	
warmup	2k steps linear	
numerical precision	bfloat16	
train steps	25k	
batch size (in images)	8k	
drop path	0.2	

A.4 Experiments Compute Resources

Pre-training used 512 A800 GPUs and took 5 days.

B Theoretical derivation details

Using the notation defined in Sec. 3, the mutual information of the output of the language model (y) and the latent representation (z) can be described as follows:

$$
\begin{aligned}
I(y; z) &= \sum_{k=1}^{N} I(y_k; z_k) \\
&= \sum_{k=1}^{N} \int p(y_k, z_k) \log \frac{p(y_k, z_k)}{p(y_k)p(z_k)} \, dy_k \, dz_k \qquad (8) \\
&= \sum_{k=1}^{N} \mathbb{E}_{x \sim p} \left[\int p(y_k \mid x) p(z_k \mid x) \log \frac{p(y_k, z_k)}{p(y_k)p(z_k)} \, dy_k \, dz_k \right] \qquad (9)
\end{aligned}
$$

We can derive Eq. 9 from Eq. 8 because once the interleaved image-text input sequence x is given, the output y_k and the latent representation z_k can be computed independently:

$$
p(y_k, z_k \mid x) = p(y_k \mid x) p(z_k \mid x)
$$

From Eq. 9, we can further decompose the mutual information into a cross-entropy component and an entropy component in two symmetric ways. One approach involves the output token y_k predicting the next latent representation z_k, along with the entropy of z_k:

$$I(y;z) = \sum_{k=1}^{N} \mathbb{E}_{x\sim p}\left[\int p(y_k \mid x)p(z_k \mid x)\log\frac{p(z_k \mid y_k)}{p(z_k)}\,dy_k\,dz_k\right] \tag{10}$$

$$= \sum_{k} \mathbb{E}_{x\sim p}\left[\delta(z_k = f_\theta(x)_k)\log P(z_k \mid y_k = g_\phi \circ f_\theta(x)_k)\right] - \int p(z_k)\log p(z_k) \tag{11}$$

The other approach considers how the latent representation z_k approximates the previous context y_k and includes the entropy of the output y_k:

$$I(y;z) = \sum_{k=1}^{N} \mathbb{E}_{x\sim p}\left[\int p(y_k \mid x)p(z_k \mid x)\log\frac{p(y_k \mid z_k)}{p(y_k)}\,dy_k\,dz_k\right] \tag{12}$$

$$= \sum_{k=1}^{N} \mathbb{E}_{x\sim p}\left[\delta(y_k = g_\phi \circ f_\theta(x)_k)\log P(y_k \mid z_k = f_\theta(x)_k)\right] - \int p(y_k)\log p(y_k) \tag{13}$$

Note that the reason for transitioning from Eq. 10 to Eq. 11 and from Eq. 12 to Eq. 13 is because, given the input sequence x, the outputs y_k and z_k are determined as follows:

$$p(y_k \mid x) = \delta\left[y_k = g_\phi \circ f_\theta(x)_k\right]$$
$$p(z_k \mid x) = \delta\left[z_k = f_\theta(x)_k\right]$$

C Supplementary Benchmark Results

We present supplementary results on classification, retrieval and captioning tasks (Tab. 10) and multi-modal benchmarks (Tab. 11).

D Broader Impacts

This work may share the common negative impacts of large-scale vision training. The data used in pre-training may contain dataset bias, and raise ethical concerns. It may also require large computational resources, which consume lots of electricity and result in increased carbon emissions.

https://doi.org/10.52202/079017-0754

https://doi.org/10.52202/079017-0754

Table 10: Supplementary transfer evaluation results on classification, retrieval and captioning tasks. * Reproducing or using open-source code. "Repro. CoCa": reproduced CoCa. Results with larger vision encoders, more training data, or private data are grayed out as reference.

Vision pre-training method	Support interleave data	Vision encoder	Pre-training data	Seen samples	Data public	IN-1k frozen	IN-1k finetune	COCO ret. frozen TR@1	COCO ret. frozen IR@1	COCO cap. finetune CIDEr
Comparison with existing vision pre-training methods on paired data										
LCL (ours)	✓	ViT-B (86M)	LAION-400M	2B	✓	75.0	-	48.3	34.3	-
*OpenCLIP		ViT-B (86M)	LAION-400M	2B	✓	75.0	-	47.2	34.2	-
*Repro. CoCa		ViT-B (86M)	LAION-400M	2B	✓	75.2	-	48.6	34.8	-
Comparison with MLLM training methods (used for vision pre-training) on interlaved data										
LCL (ours)	✓	ViT-B (86M)	MMC4	2B	✓	75.2	-	48.5	34.5	-
*BEiT3	✓	ViT-B (86M)	MMC4	2B	✓	73.3	-	45.1	30.6	-
*OpenFlamingo	✓	ViT-B (86M)	MMC4	2B	✓	24.0	-	10.6	5.6	-
*Emu	✓	ViT-B (86M)	MMC4	2B	✓	5.7	-	2.3	1.4	-
Comparison with existing pre-trained checkpoints or reported results										
LCL (ours)	✓	ViT-L (304M)	LAION-400M	13B	✓	82.2	86.1	59.6	46.2	129.7
LCL (ours)	✓	ViT-L (304M)	LAION-400M + MMC4	13B	✓	82.0	86.1	60.0	46.0	130.9
OpenCLIP		ViT-L (304M)	LAION-400M	13B	✓	82.1	86.2	59.5	46.0	128.9
OpenAI CLIP		ViT-L (304M)	WIT-400M	13B		83.7	87.4	61.7	48.2	132.7
CoCa		ViT-G (1B)	JFT-3B + ALIGN	33B		90.6	91.0	-	-	143.6

Table 11: Supplementary results of multi-modal benchmarks. † Training data of the teacher model is included. Approaches are grayed out as reference if they use stronger vision encoders or LLM, more training data, or private data.

Vision pre-training method	Vision encoder	Pre-training samples	Pre-training data public	MLLM method	LLM size	MLLM training samples	MLLM data public	VQAv2	GQA	VizWiz	SQA	POPE	MME	MMB	$SEED_I$
Our transfer evaluation															
LCL (ours)	ViT-L (304M)	13B	✓	LLaVA-1.5	7B	1.2M	✓	70.7	57.4	41.4	69.7	81.7	1291.7	55.3	56.5
OpenCLIP	ViT-L (304M)	13B	✓	LLaVA-1.5	7B	1.2M	✓	68.7	57.0	39.5	69.0	81.8	1266.2	54.2	55.8
OpenAI CLIP	ViT-L (304M)	13B		LLaVA-1.5	7B	1.2M	✓	77.1	61.7	44.4	71.1	84.6	1486.9	65.1	64.6
Advanced MLLMs as reference															
OpenAI CLIP	ViT-L-336 (304M)	13B		LLaVA-1.5	7B	1.2M	✓	78.5	62.0	50.0	66.8	85.9	1510.7	64.3	65.4
OpenAI CLIP	ViT-L-336 (304M)	13B		LLaVA-NEXT	7B	1.3M		81.8	64.2	57.6	70.1	86.5	1519	67.4	70.2
OpenAI CLIP	ViT-L (304M)	13B		InternLM-XC2	7B	∼20M		-	-	-	-	-	1712	79.6	75.9
OpenCLIP	ViT-L (304M)	34B		Qwen-VL-Chat	7B	∼1.5B		78.2	57.5	38.9	-	-	1487.5	-	-
EVA-CLIP	ViT-G (1B)	∼28B†		Emu-I	13B	∼83M		62.0	46.0	38.3	-	-	-	-	-
EVA-CLIP	ViT-E (4.4B)	∼28B†		Emu2-Chat	33B	∼160M		84.9	65.1	54.9	-	-	-	-	-
EVA-CLIP	ViT-E (4.4B)	∼28B†		CogVLM-Chat	7B	∼1.5B		82.3	-	-	91.2	87.9	-	77.6	-
InternViT	ViT-6B	∼29B		InternVL	7B	∼1.0B		79.3	62.9	52.5	-	86.4	1525.1	-	-
InternViT	ViT-6B	∼29B		InternVL 1.5	20B	∼25M		-	65.7	63.5	94.0	88.3	1637	82.2	76.0

https://doi.org/10.52202/079017-0754

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please see Sec 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please see Sec 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Please see Sec 3.1.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Details of pre-training and evaluation can be found in 4.1 and Appendix A

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

https://doi.org/10.52202/079017-0754

Answer: [Yes]

Justification: The code has been released at https://github.com/OpenGVLab/LCL. Only pulic datasets have been used in our research.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Please see Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Most experiments have stable results with little variance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please see Appendix D.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

https://doi.org/10.52202/079017-0754

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release a novel dataset or model.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original paper each time a new asset appears.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-0754