Modeling Caption Diversity in Contrastive Vision-Language Pretraining

Samuel Lavoie [1 2] **Polina Kirichenko** [* 1 3] **Mark Ibrahim** [* 1] **Mahmoud Assran** [1] **Andrew Gordon Wilson** [3] **Aaron Courville** [2] **Nicolas Ballas** [1]

Abstract

There are a thousand ways to caption an image. Contrastive Language Pretraining (CLIP) on the other hand, works by mapping an image and its caption to a single vector—limiting how well CLIP-like models can represent the diverse ways to describe an image. In this work, we introduce Llip, Latent Language Image Pretraining, which models the diversity of captions that could match an image. Llip's vision encoder outputs a set of visual features that are mixed into a final representation by conditioning on information derived from the text. We show that Llip outperforms non-contextualized baselines like CLIP and SigLIP on a variety of tasks even with large-scale encoders. Llip improves zero-shot classification by an average of 2.9% zero-shot classification benchmarks with a ViT-G/14 encoder. Specifically, Llip attains a zero-shot top-1 accuracy of 83.5% on ImageNet outperforming a similarly sized CLIP by 1.4%. We also demonstrate improvement on zero-shot retrieval on MS-COCO by 6.0%. We provide a comprehensive analysis of the components introduced by the method and demonstrate that Llip leads to richer visual representations.

1. Introduction

Contrastive Language-Image Pre-training (CLIP; Radford et al. (2021)) combined with a large-scale weakly supervised dataset has become the standard Visual Language Pretraining (VLP) approach to learn visual representation (Li et al., 2021; 2023e; Sun et al., 2023; Zhai et al., 2023; Xu et al., 2023). Due to its generality, CLIP representations are now used for many downstream tasks such as zero-shot classification (Radford et al., 2021), image generation (Ramesh et al., 2021) and visual question answering (Li et al., 2023b; Moon et al., 2023).

At its core, CLIP aims to learn an image representation that is invariant to the caption diversity (see Figure 1a). CLIP uses a visual encoder and a text encoder to independently map visual and text inputs into a common representation space. The joint encoders are trained with a contrastive objective that maximizes the similarity of representations extracted from the same image-text pair while pushing away the representations from other examples (Radford et al., 2021). This training criterion encourages the representation of an image to exactly match the representation of its corresponding text description. Further, if different text descriptions are associated with an image, CLIP contrastive objective will push both text representations toward the same visual representation.

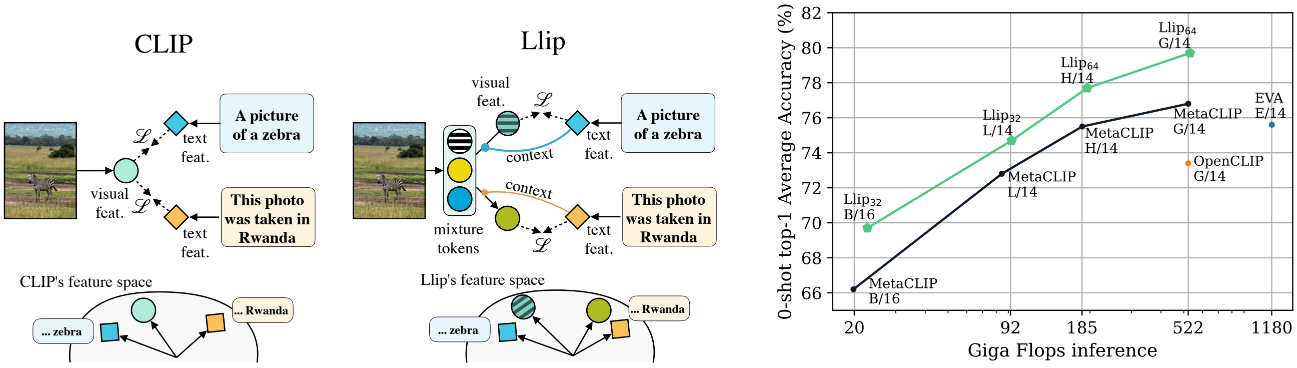
Yet, there is an information imbalance between the visual and text modality as visual content is often more rich than its text description (Foucault, 1990). Multiple diverse text captions can be equally valid descriptions of a given image, each one focusing on a different visual aspect. For example, depending on context, someone could describe the animal from the image shown in Figure 1a while another person could instead highlight the location where the picture was taken. Both are valid descriptions of the image and, arguably, different descriptions may capture different visual properties of the image. A training objective of a vision-language model should therefore aim at capturing the diversity of possible text descriptions to model the richness of the visual input.

In this work, we propose to explicitly model the fact that many different captions, and therefore representations, are plausible for a given image. To enable the prediction of different representations from a fixed image, we implement the image to text representation function as a one-to-many mapping. Conceptually, we augment our visual encoder with a latent variable that captures contextual information. Given this extra conditioning, our visual encoder can output different representations for different contexts. In our approach, the contextual latent is inferred directly from the target caption, which is then used to modulate the visual representation.

Specifically, our visual encoder is implemented by a vi-

[*] Equal contribution [1] FAIR at Meta [2] Mila, Université de Montréal [3] New York University. Correspondence to: Samuel Lavoie <samuel.lavoie.m@gmail.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.



(a) CLIP and Llip representations

(b) Averaged 0-shot classification accuracy

Figure 1: **We propose Llip, Latent Language Image Pretraining, to model the diversity of matching captions for a given image.** **(a)** Conceptual visualization of CLIP (left) and Llip (right) architectures. CLIP independently encodes visual features (shown in circles) and text features (shown in squares) which are pulled closer together by maximizing the cosine similarity objective $\mathcal{L}$. The single image feature vector of CLIP has to compromise between all matching text features (illustrated in the feature manifold at the bottom of the Figure). Llip outputs a set of *visual mixture tokens* which are combined into a final visual feature vector conditioned on *the context* derived from the caption. Llip's visual representations can more accurately represent each caption. **(b)** Zero-shot top-1 transfer accuracy averaged over 22 established classification benchmarks (see section 6.1) against Giga FLOPs for inference (estimated on the ImageNet zero-shot classification task) for encoders of various sizes. Llip outperforms the Visual Language Pretraining baselines. Llip was trained on the same data as MetaCLIP (Xu et al., 2023).

sual transformer that outputs K learnable mixture tokens in addition to the visual tokens. The goal of the mixture tokens is to capture the different visual aspects of an input. We then make use of a cross-attention mechanism that infers the mixture token weights as a function of the text caption. The weighted mixture defines our contextual representation that is contrasted with text representations. We show that this simple modification of CLIP leads to significant improvement of the visual representation quality as illustrated in Figure 1b as well as a more rich visual representation (see Figure 5). We refer to our approach as Latent Language Image Pre-training (Llip).

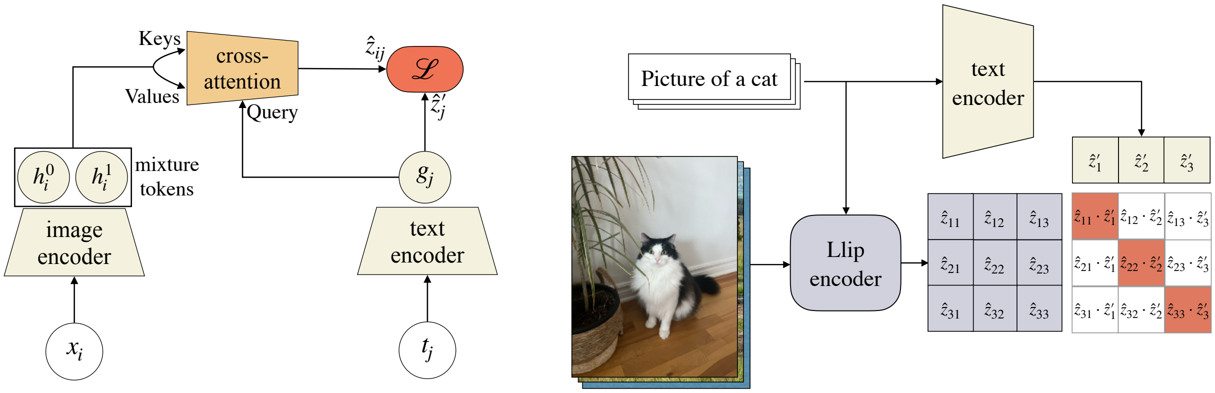
To demonstrate the value of our approach, we pretrain a family of vision transformer (ViT) encoders (Dosovitskiy et al., 2020) on the recent MetaCLIP (Xu et al., 2023) dataset and compare our approach on various zero-shot classification and text retrieval tasks. Through an empirical evaluation and control experiments we found that:

- On zero-shot transfer classification, Llip consistently outperforms CLIP pretraining for architecture of similar size on a large set of benchmarks. In particular, a VIT-G/14 encoder trained with Llip achieves a top-1 accuracy of 83.5% on the ImageNet 0-shot task outperforming a VIT-G/14 trained with CLIP by 1.4%.
- On zero-shot image-text and text-image retrieval, Llip consistently outperforms CLIP pretraining on COCO by 6.0% image-to-text retrieval.

2. Related work

Invariant representation. Invariance-based representation learning such as contrastive approaches aims at learning encoders that map two related inputs to the same point in representation space. This paradigm is commonly used in self-supervised learning (SSL) using a joint-embedding architecture (Bromley et al., 1993) where the two related inputs are two transformations of the same image (Purushwalkam & Gupta, 2020; Misra & van der Maaten, 2020; Chen et al., 2020a). In this case, the goal is to learn an invariant representation to a set of predefined image transformations that preserve the semantic content of the images (Chen et al., 2020a; Assran et al., 2022; Purushwalkam & Gupta, 2020; Misra & van der Maaten, 2020; Chen et al., 2020a; Oquab et al., 2023). While SSL methods can choose which invariance to promote through the choice of the transformations, it is not the case in vision-language pretraining as the two inputs of the encoders are from different modalities, i.e. an image and its text description. We hypothesize that enforcing invariance between image and text is not a desirable training objective as many text descriptions, capturing different visual aspects, could correspond to a given image.

Predictive representation. Another line of works in SSL learns representation without relying on invariant loss with the use of a joint-embedding predictive architectures (JEPA) (LeCun, 2022; Baevski et al., 2022; Assran et al., 2023; Bardes et al., 2024). Given a pair of related inputs


Keys
Values
cross-attention
$\hat{z}_{ij}$
$\mathscr{L}$
$\hat{z}'_j$
Query
h_i^0
h_i^1
mixture tokens
g_j
image encoder
text encoder
x_i
t_j
Picture of a cat
text encoder
$\hat{z}'_1$ $\hat{z}'_2$ $\hat{z}'_3$
Llip encoder
$\hat{z}_{11}$ $\hat{z}_{12}$ $\hat{z}_{13}$
$\hat{z}_{21}$ $\hat{z}_{22}$ $\hat{z}_{23}$
$\hat{z}_{31}$ $\hat{z}_{32}$ $\hat{z}_{33}$
$\hat{z}_{11} \cdot \hat{z}'_1$ $\hat{z}_{12} \cdot \hat{z}'_2$ $\hat{z}_{13} \cdot \hat{z}'_3$
$\hat{z}_{21} \cdot \hat{z}'_1$ $\hat{z}_{22} \cdot \hat{z}'_2$ $\hat{z}_{23} \cdot \hat{z}'_3$
$\hat{z}_{31} \cdot \hat{z}'_1$ $\hat{z}_{32} \cdot \hat{z}'_2$ $\hat{z}_{33} \cdot \hat{z}'_3$

(a) Llip encodes an image contextualized on the text features to compute the objective.

(b) Training Llip requires encoding an image with the target text caption.

Figure 2: **Summary of the method Llip.** (a) Schema of Llip's computation of the loss. An image encoder outputs K *mixture tokens* ($K = 2$ in the schema). The mixture tokens are given to a cross-attention module as keys and values along with the text encoding that is given as the query. The visual representation to be contrasted with the text target is conditioned on the text itself, allowing the model to produce a different visual representation depending on the caption. (b) Llip uses a contrastive objective and requires encoding the visual representation with the text targets to compute the loss.

x and t, JEPA approaches learn by predicting the representation of t from x conditioned on a context variable that indicates the transformation between x and t. In practice, this idea has been explored in mask-modeling formulation where the conditioning indicates the position of t (Baevski et al., 2022; Assran et al., 2023). Our approach Llip uses a similar learning principle in the context of vision-language pretraining. Our goal is to predict a text representation from the image input (see Figure 2a). One key difference with previous works is that we don't have a direct access to the conditioning variable which specifies the relative transformation from an image to its caption, Llip has to infer it using the text description.

Vision-Language Pretraining. A wide variety of prior works explored vision-language pretraining. Jia et al. (2021); Ilharco et al. (2021); Li et al. (2023d); Sun et al. (2023); Zhai et al. (2023); Fini et al. (2023); Mu et al. (2021) propose alternative contrastive-based Vision-Language Pretraining methods. Some VLP methods incorporate frozen feature extractors for image or text encoders (Zhai et al., 2022; Li et al., 2023c; Moayeri et al., 2023). Other approaches use instruction tuning (Liu et al., 2023), context (Zhou et al., 2022), and grounding objectives (Zhang et al., 2021; Li et al., 2022b; Dou et al., 2022) that require additional training data for supervision. Gao et al. (2022); Desai et al. (2024) tackle the lack of a one-to-one-correspondence between web-crawled images and captions by incorporating a hierarchical loss. All these prior works encourage invariance between image and text. Beyond contrastive pretraining, Wang et al. (2022b;a); Yu et al. (2022); Li et al. (2022a; 2023a); Dou et al. (2022) incorporate a decoder with a captioning loss into vision-language models in addition to the contrastive objective. Chen et al. (2020b); Li et al. (2021; 2020; 2022a) among others use an early or hybrid fusion of visual and text features using vision-grounded text encoder, i.e. cross-attention layers in the text encoder that attend to the output image patch tokens, which improves performance on downstream tasks but comes at a significantly increased computation cost. In our work we instead only apply a cross-attention operation to the output of vision and text encoders, and use it to mix the final visual representation vector from the mixing tokens and context inferred from the caption. In general, our approach is different from previous works in that it learns to model the diverse captions for an image solely with a contrastive objective.

3. Latent Language Image Pre-training

This section describes our proposed method: Latent Language Image Pretraining. Llip learns to output a visual representation that is conditioned on a text caption. Thus, an image have a different representation depending on the caption considered during the inference. Our approach relies on two architectural components (see Figure 2): a visual encoder that outputs K visual mixtures components, and a cross-attention module that selects how to weight the different mixture components based on the text representation.

Visual mixture tokens. The image encoder is parameterized as a Vision Transformer (ViT) (Dosovitskiy et al., 2020) which processes K learnable tokens along with each patch of the image (Darcet et al., 2023). Those learnable tokens are referred as the *visual mixture tokens*. The parameterization of our text encoder follows the CLIP's text encoder (Radford et al., 2021) and outputs a single vector representation.

Contextualization. Llip conditions the visual representation using the text representation through a multi-head cross-attention mechanism.

Let (x_i, t_i) be an image and a text caption from a dataset. We assume that x_i and t_j are a positive pair if $i = j$. Otherwise, they are a negative pair. An image encoder $x_i \mapsto \boldsymbol{h}_i$ maps an image to K visual mixture tokens $\boldsymbol{h}_i$ with h_i^k for $k \in [K]$ being the k^{th} mixture tokens. A text encoder $t_j \mapsto g_j$ maps a caption to a text feature vector.

We denote the index of each head of a multi-head cross-attention module as $m \in [M]$. The cross-attention queries are a projection of the text representation g_j: $\mathcal{Q}_j^m := g_j \cdot W_{\mathcal{Q}}^m$. The cross-attention keys and values are the projections of the visual mixture tokens: $\mathcal{K}_i^{mk} := h_i^k \cdot W_{\mathcal{K}}^{mk}$ and $\mathcal{V}_i^{mk} := h_i^k \cdot W_{\mathcal{V}}^{mk}$. The keys, queries and values of the attention are all vectors in $\mathbb{R}^{d/m}$ as defined in Vaswani et al. (2023). The mixing weights for head m are defined as:

$$\Phi_{ij}^m := \sigma_\tau((\mathcal{Q}_j^m \cdot \mathcal{K}_i^{mk})_{k=1}^K), \tag{1}$$

with σ_τ being a softmax with temperature τ computed over the K mixture tokens: $\sigma_\tau(z) := \frac{e^{z_k/\tau}}{\sum_{i=1}^K e^{z_i/\tau}}$ $\forall k \in [K]$. From the mixing weights and $\mathcal{V}$, we compute the contextualized visual representation:

$$z_{ij} := \text{Concat}\left(\left(\sum_{k=1}^K \Phi_{ij}^{mk} \cdot \mathcal{V}_{ij}^{mk}\right)_{m=1}^M\right) \cdot W_{\mathcal{O}}, \tag{2}$$

where $W_{\mathcal{O}}$ is a learnable projection matrix in $\mathbb{R}^{d \times d}$.

Similarly we project the text representation $z'_j := g_j^t \cdot W_T$ where W_T is learnable projection matrix of the text features. Both representation are normalized as previously done in CLIP when computing the objective function: $\hat{z}_{ij} = \frac{z_{ij}}{||z_{ij}||_2}$ and $\hat{z}'_j = \frac{z'_j}{||z'_j||_2}$.

Pretraining. For pretraining, we consider the SigLIP (Zhai et al., 2023) objective due to its memory efficiency. We modify SigLIP's objective using our contextualized visual representation and propose the following loss:

$$\begin{aligned}\mathcal{L}_{\text{Llip}} := \frac{1}{N}\sum_{i=1}^{N} \log \frac{1}{1 + e^{(-a\hat{z}_{ii} \cdot \hat{z}'_i + b)}} + \\ \frac{1}{N}\sum_{i=1}^{N} \sum_{j=1; i \neq j}^{N} \log \frac{1}{1 + e^{(a\hat{z}_{ij} \cdot \hat{z}'_j - b)}},\end{aligned} \tag{3}$$

where a and b are learnable parameters, N is the size of the mini-batch, $\hat{z}'_j$ is the text representation obtained from caption j and $\hat{z}_{ij}$ is the visual representation obtained from mixing the visual mixture tokens of image i with the text features of caption j.

Avoiding a shortcut solution. Contextualizing the visual features with the target caption can introduce a shortcut solution: the network ignores x_i and solely relies on t_i to minimize its objective. The negative samples of the contrastive objective in equation 3 prevent that shortcut solution. While, the caption t_i is a positive caption for x_i, the same caption is also a negative caption for a different sample x_j. Therefore, relying only on t_i is not a valid solution because the objective also minimizes the similarity for pairs of negative samples, i.e. it pushes away $\hat{z}_{ji}$ from $\hat{z}_j$.

Inference. The final visual representation depends on a caption. Consequently each image has to be encoded with all target captions as illustrated in Figure 2b, both for pretraining and zero-shot evaluation. Fortunately, the fusion of the image and text is lightweight as it occurs in the output layer. The additional compute and memory cost is constant for a fixed number of mixture tokens K as we scale up the size of the encoder (See Figure 8a).

Inference for zero-shot classification in Llip is analogous to CLIP's implementation. For a given image x_i, we have C possible caption labels $t_j, j \in [C]$. We encode each image x_i with each caption label t_j obtaining contextualized visual features z_{ij}. Then we compute the cosine similarity between the normalized visual features $\hat{z}_{ij}$ and text features $\hat{z}'_j$, and define the predicted label as the one with the highest cosine similarity between the contextualized image features and the text features.

4. Experimental Setup

Our empirical analysis over the next sections has three main objectives. First, we aim to demonstrate the contribution of each modification added by Llip via controlled experiments. Second, we illustrate the value of Llip in comparison to other contrastive VLP methods on a set of standard zero-shot benchmarks commonly used in the literature. Finally, we provide an comprehensive analysis of Llip representations and hyper-parameters. Before discussing our results, we describe our experimental setup.

We perform our experiments on 5 models: ViT-B/32, ViT-B/16, ViT-L/14, ViT-H/14 and ViT-G/14. ViT-B/32 stands for a base Vision Transformer with image patch of size 32 and ViT-L/14 is a large Vision Transformer with patch of size 14 (see Dosovitskiy et al. (2020) for implementation details). To capture the visual variability in images, our method appends K additional learnable tokens to the input sequence of transformers, similarly to Darcet et al. (2023). We refer to those extra tokens as mixture tokens and we denote the model with K mixture tokens by Llip_K. For all of our experiments, we crop and resize images to 224×224.

We pre-train our models with the AdamW optimizer (Kingma & Ba, 2017; Loshchilov & Hutter, 2017) with $\beta_2 = 0.95$ as done by Zhai et al. (2023) to stabilize the pre-training. We use a learnable scale parameter a along with a learnable bias b for our objective following the initialization of Zhai et al. (2023). Otherwise, all other training decisions closely follow the ones used by Radford et al. (2021); Xu et al. (2023). For all of the Llip experiments, we fix $M = 8$ the number of heads in the cross-attention. Unless mentioned otherwise, the cross-attention's temperature $\tau = 5$.

Our models were trained on the Common Crawl data curated using the methodology presented in Xu et al. (2023). We use a dataset of 2.5B image-text pairs collected using the same parameters that was used in Xu et al. (2023). As done in Radford et al. (2021); Xu et al. (2023) we pre-train our model for a total amount of 12.8B pairs of image-text seen with a batch size of 32,768.

To increase the training efficiency, we leverage compilation and mixed-precision in PyTorch (Paszke et al., 2019). We use gradient checkpointing for computing the activations of the visual representations to reduce the memory during pre-training. The ViT-B and ViT-L models were trained on 128 V100 and A100 respectively. The larger models were trained on 256 A100 80GB GPUs.

5. From SigLIP to Llip

To assess the impact of the contextualization of Llip, we explore how the performance evolves when gradually modifying an existing SigLIP baseline toward Llip. Our starting baseline SigLIP pre-training with a ViT-B/32 and the MetaCLIP dataset. We introduce three intermediate baselines – each corresponding to an intervention on the previous baseline – that gradually interpolate between SigLIP and Llip in the way the visual representation is computed. We present their respective performances on ImageNet zero-shot top-1 accuracy in Figure 3.

SigLIP. We reproduce SigLIP pre-training with our setup. The zero-shot accuracy on ImageNet is similar to the accuracy of 67.6 reported by MetaCLIP (Xu et al., 2023).

+ Register. We increase the amount of learned tokens from 1 to 64 in SigLIP, but only use the first learned token to compute SigLIP objective as done in Darcet et al. (2023) (they refer to additional tokens as registers). This procedure does not improve the ImageNet top-1 accuracy.

+ Average. Next, we explore the effect of tokens mixing. We compute equal-weighted average of all of the 64 learned tokens and use the resulting vector to compute the objective. We find that averaging the learned tokens leads to a significant improvement over the baseline. Adding extra learned tokens and uniform mixing is an effective method to improve VLP.

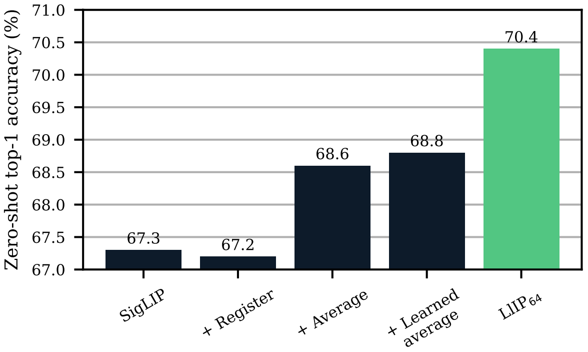


Figure 3: **Decomposing the effects of Llip's ingredients.** Ablation of the added components of Llip compared to SigLIP and their effect on zero-shot ImageNet transfer accuracy. Every models are trained with a ViT-B/32. From left to right, we evaluate: 1) Re-implemented SigLIP baseline, 2) adding additional 63 mixture tokens (+Registers (Darcet et al., 2023)) which are not used in the final representation, 3) using uniform mixing of the learnable tokens (+Average), 4) non-uniform mixing of the tokens (+Learned average), 5) context-conditional mixing of the tokens (Llip_{64}). Conditioning the mixing weights of the tokens on the text feature achieves the best performance.

+ Learned Average. We introduce non-uniform mixing to aggregate the mixture tokens. We apply a cross-attention operation as described in equation 2 except the query is a learned vector shared across all samples instead of the text caption. We don't find a significant difference between uniform and non-uniform mixing of the learned tokens.

Llip. Finally, we contrast the aforementioned baselines with Llip where the mixing weights now depend on the text features, i.e. the query token for the cross attention is a function of the text representation. Llip shows significant improvement over the average baseline in zero-shot Top-1 ImageNet accuracy.

We find that strong performance of Llip comes from mixing visual features conditioned on the text features.

6. Zero-shot Evaluations

In this section, we evaluate the performance of Llip on zero-shot classification and retrievals benchmarks. We first present an apples-to-apples comparison between CLIP, SigLIP and Llip for various backbone sizes. We train all of the models with the MetaCLIP dataset and

Table 1: **Zero-shot classification benchmarks when pretraining on the MetaCLIP dataset** on ViT-B/32, ViT-B/16, ViT-L/14, ViT-H/14 and ViT-G/14. We compare Llip to CLIP and SigLIP for several backbones with different scales. We pre-train all the models with the MetaCLIP dataset and use the same pre-training recipe. Llip outperforms MetaCLIP across most benchmarks. *: Denotes that we reproduced the baseline with our setup. MetaCLIP numbers are reported from: [1]: (Xu et al., 2023).

		Standard vision														Fine-grained					Other		
	Average	ImageNet	Food-101	CIFAR10	CIFAR100	CUB	SUN397	Cars	Aircraft	Pets	Caltech-101	Flowers	MNIST	STL-10	GTSRB	DTD	EuroSAT	RESISC45	PCAM	Country211	KITTI	UCF101	MIT-States
ViT-B/32																							
MetaCLIP[1]	62.8	67.6	82.7	95.2	77.7	67.8	66.8	77.4	27.0	90.9	92.8	69.9	42.7	96.3	39.2	**58.9**	51.1	66.3	50.0	17.7	29.3	67.5	47.6
SigLIP*	63.5	67.3	81.8	94.8	77.1	68.9	66.5	78.7	29.0	88.9	**93.0**	70.3	41.9	96.8	52.3	58.8	47.4	64.7	54.8	17.0	30.9	69.5	46.9
Llip$_{64}$	**67.5**	**70.4**	**84.1**	**95.5**	**80.8**	**71.5**	**68.6**	**82.2**	**34.9**	**92.3**	92.9	**74.8**	**66.3**	**97.5**	**53.6**	58.8	**49.9**	**67.5**	**64.5**	**20.7**	**37.8**	**71.6**	**48.5**
ViT-B/16																							
MetaCLIP[1]	66.2	72.1	88.3	95.7	79.0	71.4	68.5	82.9	30.3	91.7	93.3	73.9	66.1	98.4	46.6	62.1	51.1	71.1	50.5	22.7	16.6	73.0	50.4
SigLIP*	67.1	72.3	88.5	**96.0**	79.0	74.1	68.5	83.5	33.8	92.2	94.2	72.5	63.3	98.5	40.8	60.3	50.1	**68.6**	**55.5**	22.0	**38.2**	74.3	50.4
Llip$_{64}$	**69.7**	**75.3**	**89.0**	95.7	**81.4**	**75.0**	**70.9**	**88.2**	**41.5**	**93.5**	**94.7**	**74.9**	**79.6**	**98.5**	**54.0**	**63.7**	**56.7**	67.6	53.1	**25.7**	24.9	**77.6**	**51.7**
ViT-L/14																							
MetaCLIP[1]	72.8	79.2	93.5	97.6	84.2	80.1	73.7	88.7	44.4	94.7	95.5	81.8	64.4	**99.3**	56.3	68.3	58.7	74.6	**66.5**	34.0	29.7	81.7	55.6
SigLIP*	73.9	79.4	93.2	97.6	84.0	**82.3**	72.0	90.7	51.9	95.5	**95.7**	**83.1**	67.4	99.2	67.3	69.2	58.0	74.4	55.6	33.3	**37.4**	82.4	55.5
Llip$_{32}$	**74.7**	**80.9**	**93.6**	**98.0**	**86.8**	81.2	**74.4**	**91.7**	**55.1**	**96.0**	95.2	81.4	**68.0**	**99.3**	**68.8**	**69.8**	**59.8**	**77.3**	54.7	**36.4**	34.8	**84.5**	**56.1**
ViT-H/14																							
MetaCLIP[1]	75.5	80.5	94.2	**98.0**	86.4	83.4	74.1	90.0	50.2	95.4	95.6	85.1	72.7	**99.4**	62.5	72.4	**66.3**	**74.6**	65.8	37.2	**38.2**	82.2	56.2
Llip$_{64}$	**77.7**	**82.7**	**95.1**	97.9	**87.2**	**86.2**	**75.0**	**92.4**	**61.3**	**96.0**	**95.8**	**86.4**	**86.6**	**99.4**	**70.8**	**72.8**	62.4	74.2	**68.6**	**41.3**	33.6	**86.2**	**57.2**
ViT-G/14																							
MetaCLIP[1]	76.8	82.1	94.9	**98.5**	88.6	84.0	74.7	90.9	52.7	96.1	95.7	**89.5**	78.1	**99.5**	61.6	72.6	**73.7**	75.5	65.6	41.5	31.0	85.6	56.6
Llip$_{64}$	**79.7**	**83.5**	**95.6**	**98.5**	**89.5**	**86.8**	**76.5**	**93.6**	**67.4**	**96.7**	**95.8**	**89.5**	**89.9**	**99.5**	**72.5**	**75.7**	70.7	**77.7**	**71.9**	**45.6**	**31.1**	**88.0**	**57.9**

we fix the hyper-parameters to the one found in prior works (Zhai et al., 2023; Xu et al., 2023). We observe that Llip consistently outperforms the baselines for every model sizes on both zero-shot classification transfer and zero-shot retrieval.

Next, we compare our approach with various baselines such as CLIP (Radford et al., 2021), OpenCLIP (Cherti et al., 2023), SigLIP (Zhai et al., 2023), MetaCLIP (Xu et al., 2023), CLIPA (Li et al., 2023d), Data Filtering Network (Fang et al., 2024) that all implement a variant of constrastive learning and EVA-CLIP (Sun et al., 2023) which combines contrastive objective with input masking.

6.1. Llip improves zero-shot performance for a fixed pre-training setup

In this subsection, we evaluate Llip and compare it to the CLIP and SigLIP contrastive approaches. All methods use the same training dataset.

We evaluate Llip on a wide variety of classification benchmarks. The classification benchmarks contain tasks on object classification (ImageNet (Recht et al., 2019), CIFAR (Krizhevsky, 2010), CUB (Li et al., 2003), Food-101 (Bossard et al., 2014), STL-10 (Coates et al., 2010), caltech-101 (Li et al., 2003), MNIST (LeCun & Cortes, 2010)), fine-grained classification (SUN397 (Xiao et al., 2010), Cars (Krause et al., 2013), Aircraft (Maji et al., 2013), Pets (Parkhi et al., 2012), Flowers (Nilsback & Zisserman, 2008), GTRSB (Stallkamp et al., 2011), Country211 (Radford et al., 2021)), non-natural images (DTD (Cimpoi et al., 2013), EuroSAT (Helber et al., 2019), RESIS45 (Cheng et al., 2017), PCAM (Ye et al., 2020)) and video classification (KITTI (Geiger et al., 2012), UCF101 (Soomro et al., 2012)) and attribute recognition (MIT-States (Isola et al., 2015)).

In Table 1 demonstrates that Llip outperforms CLIP and SigLIP when controlling for the training data distribution. On a ViT-B/32, Llip outperforms SigLIP by 4.7% in average. On a ViT-G/14, Llip outperforms MetaCLIP by 2.9% in average. Table 2 also shows that Llip outperforms CLIP and SigLIP on the Flickr30k and MSCOCO zero-shot retrieval tasks. Llip outperforms a CLIP based model on MSCOCO text retrieval by 4% with a ViT-B/16 and 6% with a ViT-G/14. Llip observes similar improvement on MSCOCO image retrieval with a gain of 4.2% with a ViT-B/16 and 4.6% with a ViT-G/14.

6.2. Llip comparision with previous contrastive pre-training baselines

We now compare Llip with previously reported numbers in the literature of contrastive visual language pre-training. While these numbers are obtained with different model architectures, training recipes and datasets, we observe that Llip is a competitive method.

ImageNet. We investigate Llip's zero-shot transfer performance on the ImageNet classification task (Russakovsky et al., 2015). We report the top-1 accuracy of Llip with a ViT-G/14 and the best reported numbers from OpenCLIP, CLIP, CLIPA-v2, SigLIP, MetaCLIP and DFN in Figure 4. Llip outperforms most previous approaches. In particular,

Table 2: **Zero-shot retrieval on Flickr30k (Young et al., 2014) and MSCOCO (Lin et al., 2014).** Comparison of zero-shot retrieval performances of Llip with the SigLIP and MetaCLIP baselines. All methods are pre-trained with the same dataset and use the same pre-training recipe. We compare both Image to Text and Text to Image retrievals. Llip demonstrate consistent gain for both MSCOCO and Flicker30k. *: Reproduced number with our setup. MetaCLIP results are reported from: [1]: (Xu et al., 2023).

	Image→Text						Text→Image					
	Flickr30K			MSCOCO			Flickr30K			MSCOCO		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ViT-B/16:												
MetaCLIP[1]	85.9	97.3	98.9	59.4	80.6	87.9	70.5	90.7	94.6	41.4	67.2	77.0
SigLIP*	85.4	97.1	98.6	59.7	82.1	89.1	69.6	90.0	94.1	42.0	67.3	77.0
Llip$_{64}$	**90.1**	**98.5**	**99.6**	**63.4**	**84.3**	**90.3**	**75.1**	**92.8**	**96.2**	**45.6**	**70.8**	**79.7**
ViT-L/14:												
MetaCLIP[1]	90.4	98.5	99.1	64.5	85.0	91.3	76.2	93.5	96.4	47.1	71.4	80.3
SigLIP*	91.5	98.1	**99.4**	65.4	85.1	91.1	76.5	94.3	96.6	48.1	72.3	80.6
Llip$_{32}$	**93.2**	**99.0**	**99.4**	**68.1**	**87.6**	**92.5**	**79.9**	**95.0**	**97.4**	**50.6**	**74.7**	**82.8**
ViT-H/14:												
MetaCLIP[1]	91.6	98.6	99.7	66.2	86.2	91.9	78.0	94.6	96.9	48.8	73.2	81.4
Llip$_{64}$	**94.0**	**99.4**	**99.9**	**71.6**	**89.3**	**94.0**	**82.8**	**96.0**	**98.0**	**53.9**	**77.0**	**84.2**
ViT-G/14:												
MetaCLIP[1]	91.2	98.7	99.7	66.7	86.6	92.3	80.0	94.5	97.0	49.6	73.8	81.9
Llip$_{64}$	**94.8**	**99.7**	**100**	**72.7**	**90.1**	**94.4**	**82.5**	**96.0**	**97.9**	**54.2**	**77.1**	**84.5**

Table 3: **Comparison of zero-shot classification.** We compare Llip (*ViT-G/14*) to the best reported number of EVA-CLIP (*ViT-E/14*), OpenCLIP (*ViT-G/14*) and MetaCLIP (*ViT-G/14*) baselines on 22 classifications tasks involving object classification (e.g. ImageNet, CIFAR), fine-grained classification (e.g. Cars, Aircraft, Flowers), non-natural images (e.g. DTD, EuroSAT, PCAM). Llip obtains the best average performance across baselines and improves the best performance in 19 out of the 22 classification tasks. We only consider baselines that reports performance on the same tasks or that provide model weights. [1]: (Sun et al., 2023); [2]: (Cherti et al., 2023); [3]: (Xu et al., 2023).

	Average	ImageNet	Food-101	CIFAR10	CIFAR100	CUB	SUN397	Cars	Aircraft	Pets	Caltech-101	Flowers	MNIST	STL-10	GTSRB	DTD	EuroSAT	RESISC45	PCAM	Country211	KITTI	UCF101	MIT-States
ViT-E/14:																							
EVA-CLIP[1]	75.6	82.0	94.9	**99.3**	**93.1**	85.8	75.1	94.6	54.1	95.8	90.5	84.5	74.7	99.0	67.7	68.2	**75.8**	75.6	63.7	35.7	12.4	83.1	56.7
ViT-G/14:																							
OpenCLIP[2]	73.5	80.1	93.1	98.2	87.5	84.4	74.5	94.5	49.7	95.2	86.4	81.5	71.6	98.5	62.5	69.0	70.0	72.6	63.6	33.8	15.6	80.5	54.5
MetaCLIP[3]	76.8	82.1	94.9	98.5	88.6	84.0	74.7	90.9	52.7	96.1	95.7	**89.5**	78.1	**99.5**	61.6	72.6	73.7	75.5	65.6	41.5	31.0	85.6	56.6
Llip$_{64}$	**79.7**	**83.5**	**95.6**	98.5	89.5	**86.8**	**76.5**	**93.6**	**67.4**	**96.7**	**95.8**	**89.5**	**89.9**	**99.5**	**72.5**	**75.7**	70.7	**77.7**	**71.9**	**45.6**	**31.1**	**88.0**	**57.9**

our method shows a gain +0.3% over SigLIP while processing 4× less samples during pre-training and a gain of 2.5% over EVA-CLIP that is pre-trained with a ViT-E/14 backbone that has 2.5× more parameters that the ViT-G/14. While DFN obtains a higher zero-shot top-1 accuracy than Llip, it is trained on a larger datasets of 5B curated samples and uses 378 instead of 224 as input image resolution. We conjecture that Llip may also benefit from higher quality data, but we leave such analysis to future works.

Closest in the setting of our work is MetaCLIP which trains a joint-embedding architecture using contrastive loss on the a similar pre-training dataset. Llip outperforms MetaCLIP VIT-G/14 by +1.4%, highlighting the benefit of modelling the caption diversity.

Other image classification tasks. To demonstrate the genericity of the learned representation with Llip, we measure performances across 22 standard zero-shot classification benchmarks that are usually reported in the literature in Table 3. We compare our approach with OpenCLIP, MetaCLIP and EVA-CLIP which all report results on the same set of tasks or release their model weights allowing us to evaluate and compare with these models. Results show that Llip obtains the best average performance across baselines. It reaches the the best performance in 19 out of the 22 classification tasks.

7. Analysis of Llip

Representation expressivity. We evaluate the expressivity of the learned visual features by computing the singular values of the covariance matrix of the visual features as done in Jing et al. (2022). This method was proposed to probe the dimensionality collapse in self-supervised pretrained methods and also measures the expressiveness of learned representations (Hua et al., 2021).

In particular, we compare SigLIP, SigLIP with learned query (see Section 5) and Llip$_{64}$. We collect the embedding vectors of 5000 samples from ImageNet's validation set ran-

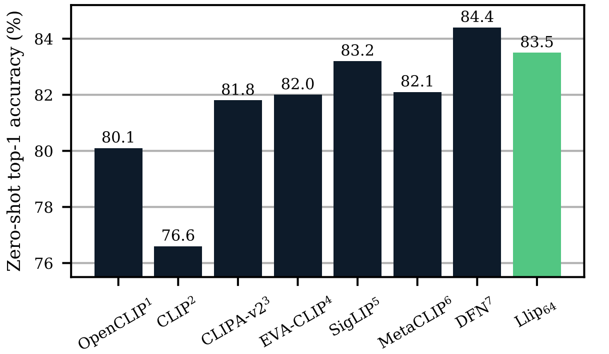


Figure 4: **ImageNet zero-shot transfer classification.** We compare a VIT-G/14 trained with Llip$_{64}$ with various vision-language baselines. We select the best reported number for every methods. Llip outperforms most of the vision-language pretraining baselines on ImageNet. Llip outperforms most of the. DFN, which is the only methods outperforming Llip, is trained on a larger datasets of 5B curated samples and use 378 instead of 224 as input image resolution. We report the imagenet performance of the baselines from: [1]: (Cherti et al., 2023); [2]: (Radford et al., 2021); [3]: (Li et al., 2023d); [4]: (Sun et al., 2023); [5]: (Zhai et al., 2023); [6]: (Xu et al., 2023); [7] (Fang et al., 2024).

domly chosen. For SigLIP with learned query and Llip, we concatenate the 64 mixture tokens along the batch dimension. Then we compute the singular value spectrum of the feature covariance matrix (Jing et al., 2022) that we plot in log scale in Figure 5. Llip show slower decay in the singular value spectrum than the two baselines which indicates a larger variability of the features.

Llip hyperparameters. Llip introduces two hyperparameters: the number of mixture tokens and the temperature of the softmax of the cross-attention module. In Figure 6 we show the result of our study on both parameters conducted with a ViT-B/32.

Number of mixtures tokens. In Figure 6a, we find that increasing the number of mixture tokens consistently improves ImageNet's top-1 accuracy without changing the model size. Moreover, as illustrated in Figure 1b, Llip's performance also scales with the model size. Llip enables three axes to scale the model: increasing the encoder's size, decreasing image patch size or increasing the number of mixture tokens.

Effect of softmax temperature. In Figure 6b, we also explore the effect of the softmax temperature. The temperature controls the sharpness of the softmax's output distribution. In each case we use the same temperature during training and inference. Higher temperatures lead to logits with higher magnitudes leading to sharper activations. Llip tends to be robust to a range of temperature values but its performance degrades for large temperatures.

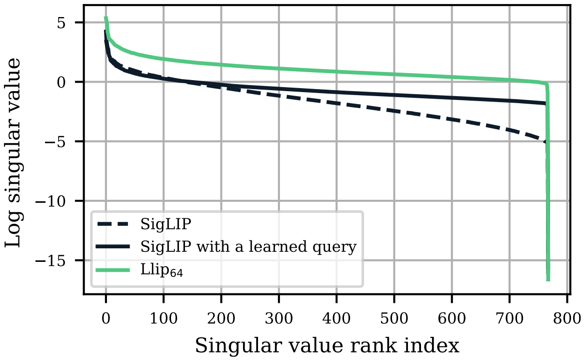
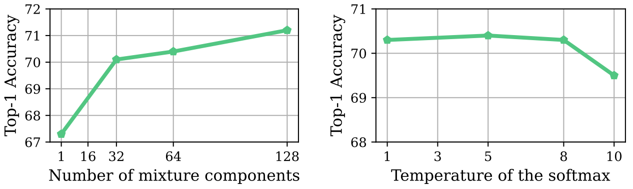


Figure 5: **Llip's representation is more expressive than the non-contextualized SigLIP baselines.** Singular value spectrum of the covariance matrix of the visual features of a ViT-B/32 using different pre-training objectives. The embedding vectors are taken at the output of the visual encoder. SigLIP with a learned query baseline adds 64 mixture tokens and learns how to average them using a cross-attention with a learnable query vector. We concatenate the 64 mixture tokens along the batch dimension for the learned query baseline and Llip. Llip show slower decay in the singular value spectrum than the two baselines which indicates a larger variability of the features.



(a) Number of mixture tokens.

(b) Attention's temperature.

Figure 6: **Analysis of Llip's hyperparameters** on downstream zero-shot top-1 ImageNet accuracy for a ViT-B/32 visual encoder. We explore the effect of the number of mixture tokens and the temperature of the softmax in the cross-attention. For (a), we set the attention temperature to 8. For (b), we fix the number of mixture tokens $K = 64$. Increasing the number of mixture tokens improves downstream performance. Llip's performance is robust to temperature values, but a large temperature leads to a degradation in accuracy.

8. Conclusion

In this work, we propose Llip – a contrastive vision-language pre-training model with contextualization of visual features to model the diversity of possible captions that could match a given image. We show that a simple approach for deriving context from the text caption and conditioning visual features leads to richer representations and better downstream zero-shot performance on a wide variety of classifications and retrieval benchmarks. Our detailed ablation studies show the benefits of each components of Llip and its robustness to hyperparameters. We hope the strength of the model on downstream tasks and its simplicity will inspire the adoption of this approach in broader scenarios.

Acknowledgements

The authors want to thank Diane Bouchacourt, Florian Bordes, Hu Xu, Pascal Vincent, Pietro Astolfi, Oscar Mañas and Micah Goldblum for insightful discussions. Aaron Courville acknowledges the Canadian Research Chair and the Cifar Canadian AI Chair for their support. Polina Kirichenko and Andrew Gordon Wilson acknowledge support from NSF HDR-2118310, CDS&E-MSS 2134216, CAREER IIS-2145492, I-DISRE 19347.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Assran, M., Caron, M., Misra, I., Bojanowski, P., Bordes, F., Vincent, P., Joulin, A., Rabbat, M., and Ballas, N. Masked siamese networks for label-efficient learning. In *European Conference on Computer Vision*, pp. 456–473. Springer, 2022.

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., and Auli, M. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 2022.

Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., and Ballas, N. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101 – Mining Discriminative Components with Random Forests. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T. (eds.), *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pp. 446–461, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10599-4. doi: 10.1007/978-3-319-10599-4_29.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

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

Cheng, G., Han, J., and Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proceedings of the IEEE*, 105(10):1865–1883, October 2017. ISSN 0018-9219, 1558-2256. doi: 10.1109/JPROC.2017.2675998.

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

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing Textures in the Wild, November 2013.

Coates, A., Lee, H., and Ng, A. Y. An Analysis of Single-Layer Networks in Unsupervised Feature Learning, 2010.

Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P. Vision transformers need registers, 2023.

Desai, K., Nickel, M., Rajpurohit, T., Johnson, J., and Vedantam, R. Hyperbolic image-text representations, 2024.

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

Dou, Z.-Y., Kamath, A., Gan, Z., Zhang, P., Wang, J., Li, L., Liu, Z., Liu, C., LeCun, Y., Peng, N., Gao, J., and Wang, L. Coarse-to-Fine Vision-Language Pre-training with Fusion in the Backbone, November 2022.

Fang, A., Jose, A. M., Jain, A., Schmidt, L., Toshev, A. T., and Shankar, V. Data filtering networks. In *The Twelfth*

International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=KAk6ngZ09F.

Fini, E., Astolfi, P., Romero-Soriano, A., Verbeek, J., and Drozdzal, M. Improved baselines for vision-language pre-training. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=a7nvXxNmdV. Featured Certification.

Foucault, M. *Les mots et les choses*. Gallimard Paris, 1990.

Gao, Y., Liu, J., Xu, Z., Zhang, J., Li, K., Ji, R., and Shen, C. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining, 2022.

Geiger, A., Lenz, P., and Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361, 2012. doi: 10.1109/CVPR.2012.6248074.

Helber, P., Bischke, B., Dengel, A., and Borth, D. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification, February 2019.

Hua, T., Wang, W., Xue, Z., Ren, S., Wang, Y., and Zhao, H. On feature decorrelation in self-supervised learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9578–9588, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.00946. URL https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00946.

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., and Schmidt, L. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773. If you use this software, please cite it as below.

Isola, P., Lim, J. J., and Adelson, E. H. Discovering states and transformations in image collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision, 2021.

Jing, L., Vincent, P., LeCun, Y., and Tian, Y. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=YevsQ05DEN7.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2017.

Krause, J., Deng, J., Stark, M., and Fei-Fei, L. Collecting a Large-Scale Dataset of Fine-Grained Cars, 2013.

Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images, 2010.

LeCun, Y. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27, 2022.

LeCun, Y. and Cortes, C. MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/, 2010. URL http://yann.lecun.com/exdb/mnist/.

Li, F.-F., Andreetto, M., and Ranzato, M. A. The Caltech-UCSD Birds-200-2011 Dataset. https://authors.library.caltech.edu/records/cvm3y-5hh21, 2003.

Li, J., Selvaraju, R. R., Gotmare, A. D., Joty, S., Xiong, C., and Hoi, S. Align before fuse: Vision and language representation learning with momentum distillation, 2021.

Li, J., Li, D., Xiong, C., and Hoi, S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, February 2022a.

Li, J., Li, D., Savarese, S., and Hoi, S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, June 2023a.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023b.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023c.

Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., Chang, K.-W., and Gao, J. Grounded language-image pre-training, 2022b.

Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., and Wang, H. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020.

Li, X., Wang, Z., and Xie, C. Clipa-v2: Scaling clip training with 81.1

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

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T. (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning, 2023.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Maji, S., Rahtu, E., Kannala, J., Blaschko, M., and Vedaldi, A. Fine-Grained Visual Classification of Aircraft, June 2013.

Misra, I. and van der Maaten, L. Self-supervised learning of pretext-invariant representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6706–6716, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society. doi: 10.1109/CVPR42600.2020.00674. URL https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00674.

Moayeri, M., Rezaei, K., Sanjabi, M., and Feizi, S. Text-to-concept (and back) via cross-model alignment, 2023.

Moon, S., Madotto, A., Lin, Z., Nagarajan, T., Smith, M., Jain, S., Yeh, C.-F., Murugesan, P., Heidari, P., Liu, Y., Srinet, K., Damavandi, B., and Kumar, A. Anymal: An efficient and scalable any-modality augmented language model, 2023.

Mu, N., Kirillov, A., Wagner, D., and Xie, S. Slip: Self-supervision meets language-image pre-training, 2021.

Nilsback, M.-E. and Zisserman, A. Automated Flower Classification over a Large Number of Classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729, Bhubaneswar, India, December 2008. IEEE. doi: 10.1109/ICVGIP.2008.47.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. V. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, June 2012. doi: 10.1109/CVPR.2012.6248092.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Purushwalkam, S. and Gupta, A. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *CoRR*, abs/2007.13916, 2020. URL https://arxiv.org/abs/2007.13916.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision, February 2021.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge, January 2015.

Soomro, K., Zamir, A. R., and Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. URL http://arxiv.org/abs/1212.0402.

Stallkamp, J., Schlipsing, M., Salmen, J., and Igel, C. The German Traffic Sign Recognition Benchmark: A multi-class classification competition. In *IEEE International Joint Conference on Neural Networks*, pp. 1453–1460, 2011.

Sun, Q., Fang, Y., Wu, L., Wang, X., and Cao, Y. Eva-clip: Improved training techniques for clip at scale, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023.

Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., and Yang, H. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework, 2022a.

Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., and Cao, Y. SimVLM: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=GUrhfTuf_3.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. SUN database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, June 2010. doi: 10.1109/CVPR.2010.5539970.

Xu, H., Xie, S., Tan, X. E., Huang, P.-Y., Howes, R., Sharma, V., Li, S.-W., Ghosh, G., Zettlemoyer, L., and Feichtenhofer, C. Demystifying CLIP Data, October 2023.

Ye, W., Yao, J., Xue, H., and Li, Y. Weakly supervised lesion localization with probabilistic-cam pooling. *ArXiv*, abs/2005.14480, 2020. URL https://api.semanticscholar.org/CorpusID:215776849.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166. URL https://aclanthology.org/Q14-1006.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models, 2022.

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., and Beyer, L. Lit: Zero-shot transfer with locked-image text tuning, 2022.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid Loss for Language Image Pre-Training, September 2023.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5579–5588, June 2021.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16816–16825, June 2022.

Appendix

A. Training setup and hyperparameters

We compare our training setup in Table 4 where we compare the training datasets, the amount of samples seen and the batch size across the methods. Llip uses the same dataset as MetaCLIP and the same batch size and amount of samples seen as MetaCLIP and CLIP. Notably, it sees less samples than the other baselines and uses a smaller dataset than SigLIP.

Table 4: Training protocol of the baselines and Llip: the dataset used, the number of samples seen during training and the batch size.

	data	samples seen	batch size
CLIP	WIT-400M	12.8B	32K
SigLIP	WebLI-10B	40B	32K
OpenCLIP	LAION-2B	39B	160K
MetaCLIP	MetaCLIP-2.5B	12.8B	32K
EVA CLIP	LAION-2B	11B+9B	144K
Llip	MetaCLIP-2.5B	12.8B	32K

The hyperparameters that we used for our method are precisely the same hyper-parameters that were used for training MetaCLIP and CLIP with the only exception of the beta2 parameter of Adam set to 0.95, the initialization of the scale and the additional bias is -10 as in SigLIP.

For zero-shot evaluation, an image has to be encoded with the target caption. Since every targets is encoded with every images and we do not know a priori which is the right target, the ground truth target cannot leak in the prediction. To reduce the compute and memory overhead in zero-shot classification, we average the text predictions and the cross-attention queries over the template axis.

B. Additional Results

B.1. Robustness

In Table 5 we show additional results on robustness benchmarks including out-of-distribution ImageNet variants across model sizes. We also show performance on geographic diversity broken down by region and model type as well as attributes from MIT States in Table 6. We find while the larger Llip model was not tuned based on the temperature parameter, when properly tuned Llip outperforms the baselines across all DollarStreet regions with a smaller encoder.

Table 5: **Robustness results on ViT-B/32, ViT-B/16 and ViT-L/14**.

	Average	Val	V2	Sketch	R	W	A
ViT-B/32							
SigLIP	57.8	67.3	59.1	56.2	76.7	58.4	28.9
Llip$_{128}$	**62.8**	71.2	62.9	60.6	82.6	62.9	36.3
ViT-B/16							
SigLIP	66.0	72.1	65.0	61.2	84.0	65.4	48.3
Llip$_{64}$	**69.7**	75.3	68.3	63.8	86.6	69.2	55.0
ViT-L/14							
MetaCLIP*	76.6	79.2	72.5	68.9	91.8	75.4	72.0
Llip$_{32}$	**79.1**	80.9	74.8	70.5	93.6	78.0	76.7

B.2. Scene and video understanding.

In Table 7, we focus specifically on scene and video understanding. We compare MetaCLIP to Llip on two scene understanding tasks (CLEVRCount, SUN397) and two video understanding tasks (KITTI, UCF101). We find the gains of Llip are more pronounced on video understanding tasks where the model obtains $+5.0\%$ on KITTI and $+2.8\%$ on UCF101.

Table 6: **Diversity across geographies.**

	Africa	Asia	Europe	Americas	Overall Top5
ViT-B/16:					
MetaCLIP	70.38	80.85	84.12	82.17	79.65
SigLIP	74.21	80.02	84.45	82.08	79.94
Llip$_{64}$	74.38	81.26	85.45	83.17	80.93
ViT-L/14:					
MetaCLIP	79.23	85.66	88.42	87.87	85.26
Llip$_{32}$	76.94	84.44	86.33	85.61	83.55

Table 7: **Scene and video understanding.** We compare MetaCLIP to Llip on two scene understanding tasks (CLEVRCount, SUN397) and two video Understanding tasks. Both models use a ViT-L/14 encoder. While Llip is competitive on both type of tasks, results show that the gain of Llip are more pronounced on video understanding tasks. MetaCLIP performance is reported from: [1]: (Xu et al., 2023).

	Scene Understanding			Video Understanding		
	CLEVR	SUN397	Avg	KITTI	UCF101	Avg
MetaCLIP[1]	**25.9**	73.6	49.8	29.6	81.6	55.6
Llip$_{32}$	25.5	**74.3**	**49.9**	**34.7**	**84.5**	**59.6**

B.3. Using image tokens in the cross-attention

While the input to Llip's vision encoder is always P image tokens and K additional visual mixture tokens, in the standard version of Llip we only use the outputs of the visual mixture tokens in the cross-attention (equation 2). In this experiment, we also included the outputs of the image patch tokens at the last layer of ViT together with the visual mixture tokens in the cross-attention (so $P + K$ tokens are used in total).

We use Llip with ViT-B/32 for which we have $P = 49$ image patch tokens, and we report results on ImageNet zero-shot classification varying the number of visual mixture tokens K in Figure 7. We train the model with temperature $\tau = 1$. We can see a similar trend as in Figure 6: the model performance increases with the higher number of the mixture tokens K.

Moreover, Llip with a smaller number of additional visual mixture tokens $K = 32$ (see Figure 6) is more effective than Llip using $P = 49$ image patch tokens and $K = 1$ mixture token (note that in the latter case the total number of tokens used in the cross-attention is higher, however, the number of additional mixture tokens used affects the performance more). We hypothesize that additional learnable tokens enable learning more expressive features leading to stronger performance.

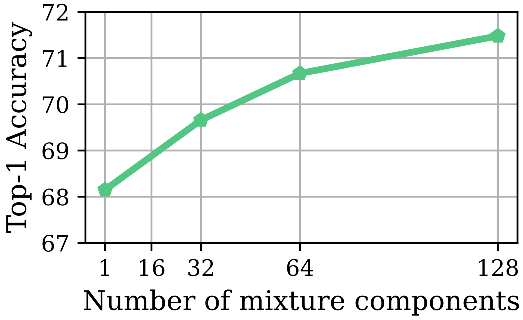
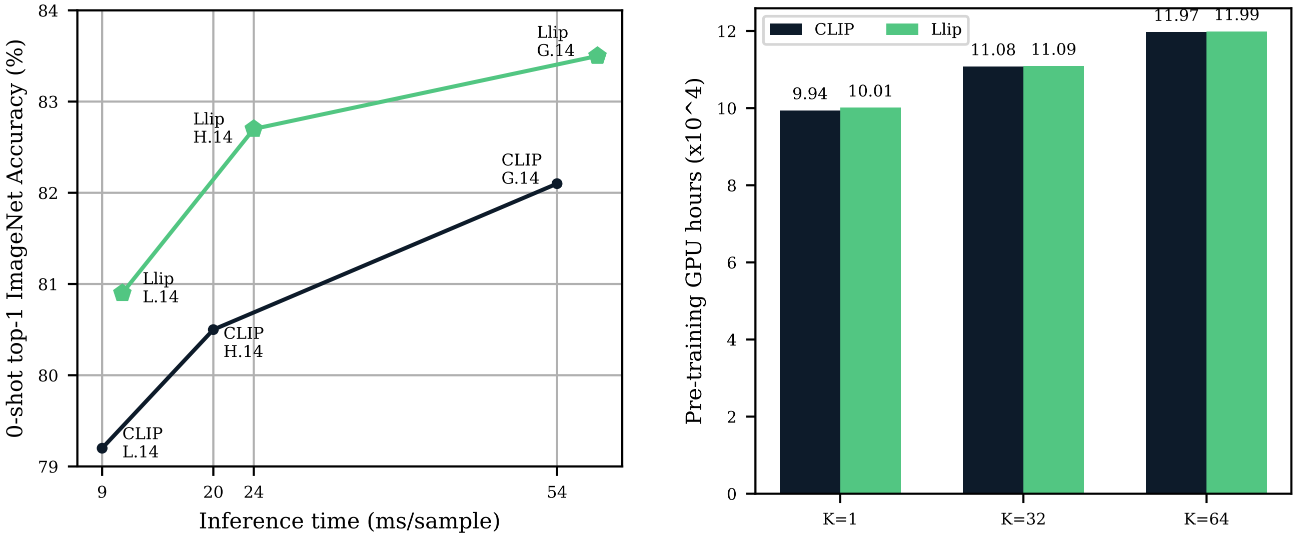


Figure 7: **Using image patch tokens together with additional visual mixture tokens in Llip.** We report zero-shot top-1 ImageNet accuracy against the number of visual mixture tokens for a ViT-B/32 visual encoder. We train Llip with temperature $\tau = 1$. Similarly to results in Figure 6, increasing the number of mixture tokens improves downstream performance.



(a) Zero-shot ImageNet accuracy top-1 accuracy against the inference time of inferring one ImageNet sample for vision encoders of various sizes.

(b) Effect of increasing the number of mixture tokens on the estimated amount of compute required for pre-training a ViT-G/14 backbone using the training recipe of (Radford et al., 2021). We find that the biggest additional cost of pre-training Llip comes from the additional mixture tokens in the vision transformer. The cost of computing the objective function is negligible.

Figure 8: Analysis of the compute overhead of using Llip's contextualization for **(a)** zero-shot inference vs. ImageNet's zero-shot transfer accuracy and **(b)** estimated pre-training GPU hours of Llip compared to CLIP.

B.4. Comparison of the compute time vs accuracy of Llip with CLIP

Inference time Figure 1b shows that the additional number of FLOPs for making an ImageNet prediction with Llip becomes marginal compared to CLIP as we scale up the encoder size. The same conclusion may be made with respect to the inference time for making an ImageNet prediction. In Figure 8a, we report the inference time for IN1K's 0-shot (1000 prompts per image) Llip's inference time is slightly higher than CLIP for the same model size, while having 1.7% improvement on 0-shot IN1K with a ViT-L/14, 2.2% with a ViT-H/14 and 1.4% with a ViT-G/14. Additionally Llip outperforms larger CLIP models while requiring a significantly lower inference time.

Pre-training GPU hours In Figure 8b, we present the amount of GPU hours that it takes for pre-training Llip and MetaCLIP for different number of mixture tokens. For estimating the amount of GPU hours, we compute the number of samples processed per hour on one A-100. We extrapolate the amount of samples processed per hour to obtain time it takes to process 12.8B samples.

While we see an increasing cost for pre-training Llip, this increase is not due to the objective of Llip. The cost of pre-training CLIP and Llip with the ViT-G/14 is almost identical when we fix the amount of mixture tokens processed by the vision transformer. Thus, the additional cost does not come from the contextualization per se, but the additional computation of the mixture tokens.