Exploring the Potential of Dense Information in Multimodal Alignment

Zhiyuan Fan[*] and **Zhihong Chen** and **Benyou Wang**[†]
The Chinese University of Hong Kong, Shenzhen
Shenzhen Research Institute of Big Data, Shenzhen, China
ambiyuan@gmail.com

Abstract

Despite the success of data augmentation in improving Contrastive Language-Image Pretraining (CLIP) model, existing methods that utilize large language model (LLM) or Segment Anything Model (SAM) to enrich the information in captions still suffer from several limitations, including insufficient detail and excessive hallucinations, ultimately resulting in compromised alignment and masking the true potential of dense information. This can lead to erroneous conclusions about CLIP's ability to handle rich data, impeding the development of more effective models. To address the limitations of existing methods, we introduce a novel pipeline that generates highly detailed, factually accurate captions for images, which facilitates in-depth analysis of the potential for dense information in multimodal alignment. Contrary to previous findings, our investigation revealed that lengthening captions boosts performance across diverse benchmarks, even surpassing the effectiveness of meticulously crafted hard negative samples. Building on these insights, DELIP is introduced, demonstrably enhancing both foundational multimodal alignment and compositional reasoning abilities. Finally, we explore strategies to expand the context window of the text encoder, unlocking the potential of richer data for CLIP and paving the way for advancements in leveraging dense information for multimodal alignment.

1 Introduction

CLIP (Radford et al., 2021) learns alignment between text and visual modalities through contrastive learning on a massive dataset of image-text pairs. Its vision encoder has been widely applied to various downstream tasks, such as image captioning (Mokady et al., 2021; Cho et al., 2023; Hessel et al., 2022), visual grounding (Xiao et al., 2024), and visual question answering (Eslami et al., 2021; Song et al., 2022). However, recent studies (Zhao et al., 2023b; Yuksekgonul et al., 2023a; Thrush et al., 2022; Lewis et al., 2023) have shown that CLIP's compositional reasoning ability is weak due to the use of noisy and simple text-image pair in the pre-training stage. This may lead to hallucinations when applying the vision encoder to downstream multimodal large language models (MLLMs). Hallucinations (Liu et al., 2024; Li et al., 2023c) refer to the model generating descriptions that are inconsistent with the input image. Since the architecture of MLLMs (Peng et al., 2023; Zhu et al., 2023; Li et al., 2023a; Liu et al., 2023b), is generally vision encoder, alignment module and large language model, if the representation obtained by the vision encoder is inaccurate or misleading, this error will be propagated.

Some studies have attempted to enhance the compositional reasoning ability of CLIP by improving the quality of captions (Doveh et al., 2023a), or constructing negative examples (Yuksekgonul et al., 2023b; Doveh et al., 2023a,b) related to objects, relationships, and attributes. One representative approach is Dense and Aligned Captions (DAC) (Doveh et al., 2023a), which enhances both the quality and density of captions and negative examples. However, the captions obtained by this method are shorter, lack details, and are more prone to hallucinations. Specifically, when using LLMs to enhance captions, they use the prompt "What can I see in a scene of {caption}?" to generate richer caption data. Obviously, this strategy of expanding captions from simple scenes will lead to the generation of details that **do not exist** in the image, i.e., hallucinations. They also proposed a method using Multiple Instance Learning (MIL) to alleviate this problem. Moreover, they only consider enhancing the density of text.

DCI (Urbanek et al., 2023) proposed a new benchmark to test the fine-grained understanding

[*] Work done during visiting student period.
[†] Corresponding author

Findings of the Association for Computational Linguistics: ACL 2024, pages 13440–13451
August 11-16, 2024 ©2024 Association for Computational Linguistics

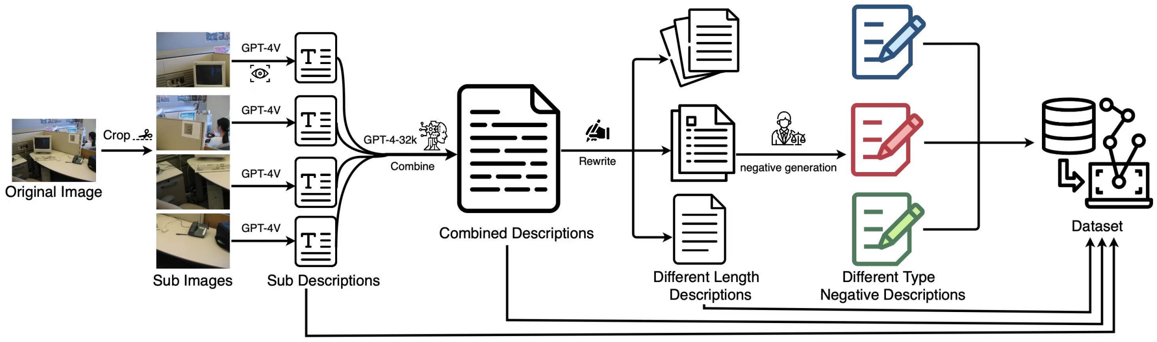

Figure 1: Data generation pipeline. After obtaining a coarse-grained global description, we crop the image into four parts to improve the relative resolution. We then request GPT-4V to generate descriptions for each part and fill the detailed information into the coarse-grained global description. Finally, we rewrite the resulting description, which contains rich details, to generate descriptions of different lengths and styles.

ability of the Vision Language Models (VLMs). They manually constructed detailed captions of each sub-crop of an image and the entire image, and tested the accuracy of the model matching each caption and subcrop. They also tested the negatives discrimination task and found that simply enhancing with hard negatives leads to a loss in the subcrop-caption matching task, demonstrating the limitations of previous work that enhanced performance just through negative examples.

In this paper, we propose DELIP, **D**ensity **E**nhanced **L**anguage **I**mage Continual **P**retraining, to analyze the impact of dense information on multimodal alignment. Specifically, we first propose a Crop-then-Merge data generation pipeline to generate detailed descriptions with fewer hallucinations, which is the basis for subsequent experiments. Then, we enhance the density of text information and image information respectively to study the impact on performance. We find that without using any tricks, we can outperform those methods that use hard negative examples on both the subcrop-caption matching task and the negatives discrimination task, which demonstrates the great potential of dense information in multimodal alignment. We then explore the strategy of expanding context windows and find that it can further improve the representation ability of the model. We also obtain some conclusions that are opposite to previous studies. Through analysis, we find that this is because the dense captions used in previous studies is either shorter or more hallucinatory. We hope that our research can help to recognize the potential of dense information and inspire future work to scale it up to a larger scale, which we call **Information Density Scaling Law**. Summarizing, our key contributions are:

1.We propose a pipeline named Crop-then-Merge for generating dense information alignment data with minimal hallucinations..

2.We conduct a detailed analysis of the impact of dense information on multimodal alignment from multiple perspectives.

3.We propose a novel multimodal alignment model, DELIP, which is based on our analysis of the impact of dense information. DELIP achieves state-of-the-art results on DCI, and also demonstrates significant improvements on zero-shot classification, text-image retrieval tasks, and compositional reasoning ability benchmarks.

2 Related Work

2.1 Multimodal Alignment

This paper delves into the alignment of text and image modalities. Established approaches like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) utilize contrastive learning from scratch, bringing together related image-text pairs in terms of their global embeddings while pushing apart unrelated ones. Filip (Yao et al., 2021) and SPARC (Bica et al., 2024) refine this strategy by performing contrastive learning at the token level, enabling more granular alignment of representations. Flamingo (Alayrac et al., 2022) leverages cross-attention to establish connections between the modalities, followed by next-token prediction training to achieve alignment. BLIP2 (Li et al., 2023a) merges three distinct loss functions – image-text matching, image-text contrastive learning, and image-grounded text generation – for comprehensive training.

Recently, numerous studies have employed alignment modules like linear projection layer (Zhu et al., 2023; Chen et al., 2023a), MLP (Liu et al., 2023b,a) or Qformer (Li et al., 2023a; Dai et al., 2023) to connect pre-trained vision encoders and large language models. Fine-tuning with visual instruction tuning data (Liu et al., 2023b; Zhao et al., 2023a; Gong et al., 2023) allows these multimodal large language model to process visual inputs. Notably, many utilize CLIP to generate visual representations or combine them with other visual models' outputs. Given the extensive use of CLIP in MLLMs and the prominent role of contrastive learning in multimodal alignment, this paper investigates the influence of dense information on such methods.

2.2 Density of Alignment Dataset

CLIP models are trained on a massive amount of image-text pairs collected from the internet. However, some works (Thrush et al., 2022; Yuksekgonul et al., 2023a) have shown that such loose and short text-image pairs contain a lot of noise, which leads to poor compositional reasoning ability of the model.

LaCLIP (Fan et al., 2023) improves the diversity of text information by rewriting the style of the original caption, but this may introduce additional hallucinations. MLLM-augmented CLIP (Liu et al., 2023c) uses multiple MLLMs to generate caption data, which improves the diversity of expression and reduces hallucinations, but still covers little visual information. DAC (Doveh et al., 2023a) uses a two-stage approach to augment the data: first using MLLM to generate the base caption, and then using Segment Anything Model (SAM) (Kirillov et al., 2023) and LLM to augment the caption respectively, to obtain dense data. However, this approach introduces a lot of hallucinations, and in order to mitigate the impact of hallucinations, they propose to use the Multiple Instance Learning to mitigate this noise. FLIP (Li et al., 2023b) borrows the idea of MAE (He et al., 2021), due to the sparsity of image information, masks 75% of the patches of the input image, and accelerates the training process without sacrificing performance, which inspires us to explore the impact of enhancing the information density of images. All of these methods use text data lengths that are much less than 77, and FLIP even sets the max token of the text encoder to 32.

In this paper, we explore the impact of continual training of CLIP models with dense text information and image information. In terms of text information, we analyze the impact of length, style, and negative samples. In terms of image information, we analyze the impact of the visual richness.

3 Approach

3.1 Dataset Generation Pipeline

We found that directly using MLLMs to generate image descriptions will miss some details, and if the model is forced to continue to output more details, the model is prone to hallucination. We think this is because of the resolution of the input image to the MLLM, which is 224*224 or 336*336, the model cannot see the specific details of the image and can only guess based on the high-level concept of the image.

We propose a **Crop-then-Merge** pipeline to generate detail-rich, low-hallucination data, as shown in Figure 1. Specifically, first, we generate a general but detail-poor description of the input image at low resolution. Then, we divide the input image into four parts and send each part to the GPT-4V (OpenAI et al., 2024) to generate a description of each sub-image, which gives each sub-image a larger relative resolution and can capture richer and more accurate details. Next, we use the sub-image generated descriptions to enhance the general description, resulting in a description that contains richer details.

Additionally, in order to study the impact of training data with different styles, lengths, and negative samples on model performance, we perform different degrees of summary on the obtained detail enhanced global description to obtain long, middle, and short descriptions, and then rewrite the descriptions to generate different styles. Finally, multiple hard-negative samples are generated for each description.

Different from the previous method of simply replacing objects, attributes, and relationships in the description, we use LLM combined with carefully designed prompts to generate each negative sample, which is more natural and coherent, conforms to normal world commonsense, and is more difficult for the text encoder to discriminate.

3.2 Loss Function

We employed the InfoNCE (van den Oord et al., 2019) loss function during our analysis of the influence of various factors on model performance. The fundamental contrastive loss function used is given

Dataset	Mean	Std
Laion Original Caption	11.39	11.64
General Description	180.91	46.77
Merged Description	579.63	67.08
Short Version Rewrite	97.09	26.21
Middle Version Rewrite	253.64	29.83

Table 1: Mean and Standard Deviation for Different Datasets Lengths.

by:

$$L = -\log\left(\frac{\mathrm{sim}(I_i, T_i)}{\mathrm{sim}(I_i, T_i) + \sum_{j=1}^{N} \mathrm{sim}(I_i, T_j)}\right)$$

where I_i denotes the i^{th} image, T_i denotes the text matching, T_j denotes all negative text samples, $\mathrm{sim}(I, T)$ denotes the similarity between image I and text T.

Within each epoch of the final training phase, we randomly select one positive sample from a pool encompassing diverse lengths and styles, while incorporating all negative samples.

4 Experiment

4.1 Experimental Setup

In order to accommodate the amount of trainable model parameters and training data, we adopted different training configurations. In the process of searching for training hyperparameters and selecting prompts, we utilized training and evaluation sets of sizes 9500 and 500, respectively. The experiments were conducted on 8 * V100 GPUs with a per device batch size of 16, in order to keep a consistence with DCI.

We first hand-wrote a set of prompts as the sampling pool, and then used a progressive optimization approach to find the optimal prompt for each description generation stage. Our evaluation set was the aggregation of data generated by these prompts, while the training set was just the result of a single prompt. Ultimately, we selected the hyperparameters and prompts with the lowest loss on the sum eval dataset. In the first stage, we used GPT-4V to generate descriptions for each sub-image. In the second stage, we used GPT-4-32k to aggregate the generated descriptions based on the global description. Since the third stage of data generation was mainly text rewriting (length, style), we found that GPT-turbo-3.5 could complete this task very well. In order to save costs, we did not use the GPT-4 interface. Since GPT-4 and GPT-3.5-turbo have been used as the target of open source models for a long time, and it is difficult to determine which open source model is the best, we used the OpenAI interface instead of deploying open source models locally.

The training process employed 8 * A100 80G servers, with a per-device batch size of 128. To mitigate catastrophic forgetting, two training methodologies were utilized: LoRA (Hu et al., 2021) and fully fine-tuning. In the LoRA training, an r value of 32, alpha of 32, and dropout rate of 0.1 were applied, specifically targeting the Q, K, and V modules. The corresponding learning rate was set to 5e-4. For full fine-tuning, a significantly lower learning rate of 5e-6 was chosen. An AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate warm-up to 3e-4 and a linear decay rate of 0.9 governed the optimization process. Training continued for 10 epochs.

During our experiments, we observed that in terms of speed, training conducted in a distributed manner showed no significant increase in speed when using LoRA compared to fully finetuning. In terms of performance, although LoRA reduced memory usage and allowed for an increased batch size, the loss associated with LoRA was significantly higher than that of fully finetuning. As a result, the test results on benchmarks were notably lower for LoRA compared to fully finetuning. Therefore, we ultimately chose to use fully finetuning for subsequent experimental analysis.

4.2 Evaluation

4.2.1 DCI

DCI (Urbanek et al., 2023) contains precise and reliable captions associated with specific parts of an image, which averages more than 1000 words. It evaluates the fine-grained understanding ability of VLMs through the task of matching each caption with its corresponding subcrop. To facilitate the evaluation of existing VLM models, they also released a version of sDCI that summarizes these detailed annotations to 77 tokens due to context limit.

4.2.2 ARO & VL-CheckList

Multiple benchmarks (Thrush et al., 2022; Zhao et al., 2023b; Yuksekgonul et al., 2023b) are proposed to evaluate the compositional reasoning ability of VLMs. In our work, to ensure consistency

	All		All Pick5		Base	All
Model	SCM	Neg	SCM	Neg	Neg	Hard Neg
CLIP Baseline	40.12%	60.63%	11.28%	24.03%	67.66%	41.29%
NegCLIP	43.35%	56.00%	13.22%	4.82%	76.69%	50.84%
BLIP	39.13%	54.02%	10.73%	5.51%	63.41%	53.23%
Flava	38.08%	47.99%	8.01%	9.82%	11.6%	45.59%
X-VLM	38.45%	53.46%	10.96%	5.10%	44.29%	52.42%
DAC$_{LLM}$	37.48%	81.75%	8.27%	37.96%	90.54%	71.29%
DAC$_{SAM}$	37.90%	84.22%	6.78%	39.91%	89.68%	73.68%
DELIP$_{pos}$	41.79%	63.54%	10.46%	19.97%	67.85%	59.23%
DELIP$_{pos-mix}$	39.28%	61.75%	10.21%	18.52%	65.43%	55.88%
DELIP$_{pos-inter}$	39.19%	60.39%	8.41%	21.81%	71.42%	57.13%
DELIP$_{bert}$	39.81%	69.56%	11.49%	28.72%	79.46%	63.13%
DELIP$_{bert-mix}$	39.60%	64.50%	11.96%	23.80%	79.46%	56.22%
DELIP$_{10k}$	43.91%	64.98%	13.41%	22.44%	78.57%	54.17%
DELIP$_{vision-rich}$	44.01%	64.72%	14.14%	22.13%	77.26%	54.73%
DELIP$_{all}$	**44.93%**	65.71%	**15.59%**	23.96%	83.93%	55.88%
DELIP$_{neg}$	44.17%	80.73%	14.98%	35.17%	89.77%	69.36%

Table 2: Test Results on the sDCI Dataset. Here, SCM refers to the subcrop-caption matching task, which measures the model's ability to understand images and text at a fine-grained level. Neg measures the model's ability to discriminate negatives, and pick 5 refers to using 5 negative examples. Our model significantly outperforms others in the SCM task and exhibits superior performance in the Neg task without the need for hard negative samples. Variants using a pretrained text encoder show considerable improvements in discriminating negative examples.

with previous research, we adopted the widely used ARO (Yuksekgonul et al., 2023b) and VL-CheckList (Zhao et al., 2023b). These benchmarks specifically target the assessment of models' compositional reasoning ability of objects, relationships, attributes and order information.

4.2.3 CLIP-Benchmark

Prior work predominantly focused on improving the compositional reasoning ability of models and solely reported results on related benchmarks. However, we discovered that these methods can potentially affect the foundational multimodal capabilities of the model. To address this, we evaluate the model's basic multimodal abilities using the CLIP-Benchmark. Specifically, text-image retrieval is employed to assess the model's cross-modal alignment capacity, while zero-shot image classification serves to evaluate its generalization and semantic comprehension abilities.

Given the widespread application of CLIP's vision encoder directly in MLLMs or in conjunction with DINO (Caron et al., 2021; Oquab et al., 2024) to provide hybrid representations (Lin et al., 2023; Jiang et al., 2024), independent evaluation of the vision encoder's visual representation capability is of paramount importance. In image-text retrieval and zero-shot retrieval, performance is influenced by both the vision encoder and the text encoder, making it challenging to attribute performance improvements solely to the vision encoder. In Section 4.5, we compare the results of training both the vision and text encoders with those of training only the text or vision encoder. We also evaluate the visual representation capability of the vision encoder through linear probing.

4.3 Text-rich Information

We used the data generation pipeline in Section 3.1 to generate data on the Laion (Schuhmann et al., 2021) dataset to analyze the impact of dense text information on model performance. We gradually increased the size of the dataset and trained the model while observing various metrics. When we found that the model's performance was saturated, we stopped generating new data to save costs. Finally, the size of the dataset we obtained was 493k.

Our study reveals that, limited by a context window of 77, dense textual information significantly boosts the model's performance on DCI benchmark, as shown in Table 2. This contradicts the DCI finding that improved performance via negatives-based training comes at the cost of decreased performance on subcrop-caption match-

Model	VL-Checklist Object	VL-Checklist Attribute	VL-Checklist Relation	ARO VG-R	ARO VG-A	ARO COCO	ARO FLICKR
CLIP	81.58	67.60	63.05	59.98	63.18	47.90	60.20
BLIP2	84.14	80.12	70.72	41.16	71.25	13.57	13.72
NegCLIP	81.35	72.24	63.53	81.00	71.00	86.00	91.00
DAC$_{LLM}$	87.30	77.27	86.41	81.28	73.91	94.47	95.68
DAC$_{SAM}$	88.50	75.83	89.75	77.16	70.50	91.22	93.88
DELIP$_{10k}$	82.17	71.16	66.08	64.27	67.39	74.30	75.21
DELIP$_{all}$	84.49	74.43	68.81	66.32	69.74	79.08	79.39
DELIP$_{neg}$	87.90	78.01	87.23	80.91	74.32	93.35	94.74

Table 3: Test results for Compositional Reasoning Ability. We compared using only 10k data, using all data, and using all data with additional hard negative samples with other methods. Note that the DAC series all used Multiple Instance Learning and hard negative samples.

ing. Notably, we demonstrate that dense information simultaneously enhances both subcrop-caption matching task and negatives discrimination task without resorting to extra hard-negative samples. This achievement surpasses the performance of NegCLIP (Yuksekgonul et al., 2023b) and other models. We speculate that this is because dense information contains more fine-grained alignment and more complex grammatical structures.

Similarly, our experimental results on various benchmarks also contradict Liu et al.'s (2023c) claims that using long descriptions will reduce model performance. We analyzed the differences between the data generated by simply using open source MLLMs and the data generated by our pipeline. We found that the data generated by the single MLLM has more hallucinations (Liu et al., 2024; Zhou et al., 2023; Liu et al., 2023c). Although the length is long, there are a lot of hallucinations in the latter half of the description, which may damages the potential of dense information.

Furthermore, we investigated the impact of scaling up the amount of dense text information on performance. Our findings demonstrate that using only 10k data points can significantly improve the model's performance on the DCI, ARO, VL-Checklist and CLIP-Benchmark, highlighting the efficiency of dense information.

While our study utilized dense descriptions, their length distribution diverged from shorter captions. Despite this, the DELIP model achieved enhanced performance in text-image retrieval, as shown in Table 4. We posit that this improvement stems from the inherent complexity and diversity of sentence structures within dense descriptions, coupled with their fine-grained detail. These characteristics align well with the simpler caption pattern typically employed in text-image retrieval tasks.

Table 5 reveals mixed results for the DELIP model across various datasets on zero shot classification task, exhibiting both improvements and declines. This inconsistency might be attributed to a mismatch between the prompt template for zero shot classification and the stylistic characteristics of our training data. In order to investigate the impact of different data generation styles on the model, we compared the effects of using data generated with a single prompt and data generated with a mixture of different prompts during the training phase. To improve computational efficiency, when using mixed data, a positive sample was randomly selected in each epoch, and the rest were used as in-batch negative samples. We found that using data with different styles leads to different improvements on different tasks, which is consistent with previous studies on the impact of training data distribution bias on model performance (Li et al., 2023b). However, despite the slight differences, the overall trend is upward, which demonstrates the robustness of dense information.

Although using dense information already outperforms methods that use hard-negative samples, we still investigated whether using negative samples can further improve the model's performance. We found that using hard negative samples can significantly enhance the model's compositional reasoning ability, but it has a certain degradation on zero-shot classification, which is consistent with the previous conclusions on using hard negative samples.

4.4 Vision-rich Information

We observed that the images in the Laion dataset are relatively simple and contain less visual infor-

	MSCOCO		Flickr30k		Flickr8k	
Model	I2T R@5	T2I R@5	I2T R@5	T2I R@5	I2T R@5	T2I R@5
CLIP Baseline	56.00	74.96	83.24	94.9	80.46	91.40
DELIP$_{pos}$	21.69	41.17	46.97	70.89	41.47	62.80
DELIP$_{pos-mix}$	47.38	61.08	75.27	87.5	69.37	81.69
DELIP$_{bert}$	22.11	33.89	41.53	56.90	39.71	54.90
DELIP$_{t}$	46.00	71.89	36.68	58.20	16.73	30.68
DELIP$_{v}$	56.93	72.18	84.25	93.30	82.02	90.49
DELIP$_{10k}$	56.55	72.93	81.49	93.50	81.49	91.50
DELIP$_{all}$	58.33	74.65	84.96	93.80	82.36	92.59

Table 4: Test results for Text-Image Retrieval.

mation. FLIP even masked out 75% of the patches to increase the batch size and accelerate training. Based on this observation, we explored the impact of using images with richer visual information on performance.

To maintain consistency, we used the same prompt and data generation pipeline to generate text descriptions. The only difference was that we used images from the SAM dataset. Due to the cost of the GPT-4V API, which is related to the pixel size of the input image, and the high resolution of the images in the SAM dataset, we only generated 10k training data based on SAM to save costs. In this paper, we mainly analyze the impact of dense visual information on performance under the same amount of data, and leave the work of scaling up data to future work.

4.5 Text-only v.s. Vision-only Tuning

Our findings suggest that fine-tuning either the text encoder or the vision encoder in isolation is insufficient for optimal performance. This limitation might stem from the inability of the text encoder to effectively represent dense text information during vision-only fine-tuning. The model, trained primarily on short and sparse caption data, might struggle with the richer and more complex nature of dense text, leading to degradation of the vision encoder's performance as measured by linear probing. Additionally, fine-tuning the text encoder alone exhibits performance losses, implying that dense text goes beyond simply increasing length and sentence complexity. It introduces fine-grained alignment between the description and the image, necessitating joint training of both encoders for optimal learning.

Based on the linear probing results presented in Table **??**, there appears to be minimal performance difference between our model and other approaches. We think the reason may be 1. the capacity limitation of the vision encoder model; 2. DELIP has enhanced the ability to understand images at a fine-grained level, but the linear probe tests the model's understanding of global information.

4.6 Scale up the Model Size

We found that increasing the capacity of the model can improve the performance of the model on DCI. We hope that our work can inspire future research to use dense information to scale up the model to billions or even larger scale, while current works (Dehghani et al., 2023; Sun et al., 2024; Chen et al., 2024) only scales up the size of the model, still using short text descriptions and sparse images.

5 Expand the Context Length

Previous works were hindered by a context window size of 77, preventing the model from effectively utilizing the information contained in longer captions. To address this limitation and harness the potential of denser text information, we explore increasing the context window to 512 tokens and analyze its impact on model performance.

5.1 Strategies

We use the following strategies to expand the context windows:

Train Position Embedding from Scratch: The length of the position embedding is the primary factor limiting the context window. Based on this, we reinitialize the position embedding with a length of 512 and then retrain the entire model. To investigate the impact of the length distribution of the text data, we compared two strategies: training only with long descriptions and training with a mixture

	FER2013		**ImageNet-1k**		**ImageNet-v2**		**MNIST**		**VOC2007**		**CIFAR-100**	
Model	Acc1	Acc5	Acc1	Acc5	Acc1	Acc5	Acc1	Acc5	Acc1	Acc5	Acc1	Acc5
CLIP	41.43	94.75	63.35	88.82	55.75	83.41	48.86	85.17	76.47	95.94	64.24	88.76
DELIP$_{position}$	40.59	88.56	17.63	37.36	15.29	34.13	36.68	77.53	40.82	80.22	25.20	51.44
DELIP$_{mix}$	42.89	92.78	29.22	55.59	25.1	50.21	30.48	82.3	51.94	81.25	35.87	65.09
DELIP$_t$	28.79	80.11	23.74	40.56	20.32	37.2	25.62	62.85	4.98	27.6	51.41	78.05
DELIP$_v$	48.76	92.74	55.04	82.59	47.93	76.23	51.16	83.45	67.50	91.15	54.02	81.86
DELIP$_{10k}$	51.28	94.78	53.64	81.61	46.61	75.14	48.34	88.03	68.97	92.01	50.48	79.29
DELIP$_{all}$	53.57	95.31	54.01	82.79	47.27	75.05	48.34	86.49	63.13	90.16	51.79	78.23

Table 5: Zero-Shot Classification Test Results. We did not find the best results by carefully designing prompt templates. We used the default ones in CLIP-Benchmark.

of long descriptions and original captions.

Position Interpolation: We use interpolation (Chen et al., 2023b) to expand the position embedding; specifically, the position embedding is interpolated to a length of 512 as the initial embedding, and then we continue to train the model with the same data settings.

Pretrained Text Encoder: Besides expanding the position embedding based on the text encoder of CLIP, another approach is to use a pretrained language model with a longer context window as the text encoder. AltClip (Chen et al., 2023c) and NLLB-Clip (Visheratin, 2023) used a pretrained multilingual text encoder for replacement, which facilitates the performance of CLIP on low-resource languages. Different from previous work, we explore the impact of using pretrained language models on processing longer text descriptions.

5.2 Analysis

We found that if we only use long descriptions to train position embeddings from scratch, the model can perform well on DCI, but the performance loss is very serious on clip-benchmark; after mixing the original caption data for training, the model's performance on zero shot classification and retrieval tasks is restored, but still lower than CLIP-base. We think this is due to the insufficient training data, we only used 493k data for training, which is lower than the data required for pre-training. Due to the limitations of API cost and computing resources (the time and memory required to train the model with a context window of 512 is much higher than 77), and the main purpose of this paper is to explore the potential of dense information rather than pre-training, we leave using dense information to train the clip model from scratch as future work.

The effect of position interpolation is not very good. We think it is because CLIP's position embedding is obtained through training rather than rotary position embeddings (Su et al., 2023; Touvron et al., 2023). The gap between these two methods of injecting position information leads to position interpolation not being very suitable for CLIP-style models. A potential high-efficiency solution is to use a text encoder with rotary position embedding to pre-train on short text, and then extend it to long text through position interpolation combined with fine-tuning.

While incorporating longer context via BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) models did improve overall results, the difference between their impacts was marginal. Interestingly, training solely on dense descriptions led to more strong performance in negative discrimination, even without hard negative samples. However, combining dense descriptions with original captions caused a performance drop in negative discrimination, while subcrop-caption matching remained unaffected. This suggests that dense descriptions have potential to enhance the model's ability to differentiate hard negatives, a crucial aspect of its representation ability.

6 Future Work

Firstly, in this paper, we use GPT-4V to generate descriptions for each image or subcrop and then use GPT-4 to merge these generated descriptions into a global description with detailed details while minimizing hallucinations. However, calling the GPT-4 API is expensive. Therefore, it is necessary to explore how to use open-source MLLMs to generate high-quality descriptions and how to use open-source LLMs to merge different descriptions together.

Secondly, we have only divided a single image into four parts, which still may overlook some details if the original image's input resolution is high and contains many details. A potential direction worth exploring is to divide the image into more parts to generate descriptions separately, which would improve the overall richness of details. However, it is essential to note how to effectively merge these individual parts without introducing additional hallucinations (e.g., if an object is incorrectly divided into multiple parts), which is a problem worth studying.

Thirdly, when merging the descriptions of each part, we explicitly marked the relative position of each part in the original image (e.g., top left, bottom left, top right, bottom right) in the prompt. However, if the image is divided into more parts, marking the position information of each part becomes difficult. A straightforward idea is to use a coordinate system for annotation. We leave this part of the exploration work for the future.

7 Conclusion

We conducted a systematic analysis of the impact of dense information on multimodal alignment. Analyzing the impact of increasing the density of text information and vision information on model performance, we found that data rich in detail and with less hallucination improved the model's fine-grained alignment and enhanced its ability to distinguish negatives. Interestingly, simply using density-enhanced training data already outperformed using carefully designed hard negative samples. We also tried various strategies to further increase context windows, finding that it further improved the model's representation ability, showcasing the great potential of dense information for enhancing multimodal alignment.

Limitations

In this paper, we utilized continual training to explore the impact of dense information on multimodal alignment. Due to the analysis of multiple variables, the total experimental time reached 21 days on 8 * A100 80G servers, despite employing base-sized models.

1. **No Pretraining for Dense Information Impact Observation:** While the configuration and time were sufficient for pre-training CLIP-base, our employed description length significantly exceeded the original captions in the Laion400m dataset. This resulted in limitations in batch size and training speed. 2. **Data Saturation via GPT-4v Only:** To optimize cost, we used GPT-4v to generate data sufficient for model saturation, without creating further variations. Scaling up the dataset alongside the model size, though not implemented due to cost constraints, was expected to improve performance. 3. **English-Only Experiments:** While a body of work exists on multilingual CLIP, our data construction and experiments exclusively utilized the English language. Investigating how to construct dense image-text pairs in low-resource languages presents a valuable avenue for future exploration.

Ethics Statement

By enhancing information density, we have improved multimodal alignment. However, hallucination remains unavoidable, which may produce outputs that do not align with facts and could mislead users. Therefore, it is necessary to manually review the model's outputs to ensure safe use in downstream applications.

Acknowledgments

This work was supported by the Shenzhen Science and Technology Program (JCYJ20220818103001002), Shenzhen Doctoral Startup Funding (RCBS20221008093330065), Tianyuan Fund for Mathematics of National Natural Science Foundation of China (NSFC) (12326608), Shenzhen Key Laboratory of Cross-Modal Cognitive Computing (grant number ZDSYS20230626091302006), and Shenzhen Stability Science Program 2023, Shenzhen Key Lab of Multi-Modal Cognitive Computing.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning.

Ioana Bica, Anastasija Ilić, Matthias Bauer, Goker Erdogan, Matko Bošnjak, Christos Kaplanis, Alexey A. Gritsenko, Matthias Minderer, Charles Blundell, Razvan Pascanu, and Jovana Mitrović. 2024. Improving fine-grained understanding in image-text pretraining.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023a. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023b. Extending context window of large language models via positional interpolation.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks.

Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Qinghong Yang, and Ledell Wu. 2023c. AltCLIP: Altering the language encoder in CLIP for extended language capabilities. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8666–8682, Toronto, Canada. Association for Computational Linguistics.

Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. 2023. Fine-grained image captioning with clip reward.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin F. Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Patrick Collier, Alexey Gritsenko, Vighnesh Birodkar, Cristina Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetić, Dustin Tran, Thomas Kipf, Mario Lučić, Xiaohua Zhai, Daniel Keysers, Jeremiah Harmsen, and Neil Houlsby. 2023. Scaling vision transformers to 22 billion parameters.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. 2023a. Dense and aligned captions (dac) promote compositional reasoning in vl models.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. 2023b. Teaching structured vision&language concepts to vision&language models.

Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. 2021. Does clip benefit visual question answering in the medical domain as much as it does in the general domain?

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2023. Improving clip training with language rewrites.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2021. Masked autoencoders are scalable vision learners.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. Clipscore: A reference-free evaluation metric for image captioning.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision.

Dongsheng Jiang, Yuchen Liu, Songlin Liu, XIAOPENG ZHANG, Jin Li, Hongkai Xiong, and Qi Tian. 2024. From CLIP to DINO: Visual encoders shout in multi-modal large language models.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything.

Martha Lewis, Nihal V. Nayak, Peilin Yu, Qinan Yu, Jack Merullo, Stephen H. Bach, and Ellie Pavlick. 2023. Does clip bind concepts? probing compositionality in large image models.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. 2023b. Scaling language-image pre-training via masking.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi

Zhang, Xuming He, Hongsheng Li, and Yu Qiao. 2023. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

Yanqing Liu, Kai Wang, Wenqi Shao, Ping Luo, Yu Qiao, Mike Zheng Shou, Kaipeng Zhang, and Yang You. 2023c. Mllms-augmented visual-language representation learning.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. Clipcap: Clip prefix for image captioning.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien

Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. Dinov2: Learning robust visual features without supervision.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs.

Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. 2022. Clip models are few-shot learners: Empirical studies on vqa and visual entailment.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. Roformer: Enhanced transformer with rotary position embedding.

Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. 2024. Eva-clip-18b: Scaling clip to 18 billion parameters.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. 2023. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding.

Alexander Visheratin. 2023. Nllb-clip – train performant multilingual image retrieval model on a budget.

Linhui Xiao, Xiaoshan Yang, Fang Peng, Ming Yan, Yaowei Wang, and Changsheng Xu. 2024. Clip-vg: Self-paced curriculum adapting of clip for visual grounding. *IEEE Transactions on Multimedia*, page 1–14.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023a. When and why vision-language models behave like bags-of-words, and what to do about it?

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023b. When and why vision-language models behave like bags-of-words, and what to do about it?

Bo Zhao, Boya Wu, Muyang He, and Tiejun Huang. 2023a. Svit: Scaling up visual instruction tuning.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2023b. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models.