Aligning Vision Language Models with Contrastive Learning

Kenan E. Ak*, Jay Mohta*, Dimitris Dimitriadis, Saurav Manchanda, Yan Xu, and Mingwei Shen

Amazon.com
{kenanea, jaymoht, dbdim, sauravma, yanxuml, mingweis}@amazon.com

Abstract. In recent years, Vision Language Models (VLMs) have achieved significant advancements due to the success of large language models. The common strategy for aligning vision and language models involves a two-step process: an alignment (or pretraining) stage and an instruction tuning stage. During the alignment stage, a projection module is trained to map image embeddings into the language space using a paired image-text dataset. In the instruction tuning stage, the model is trained to answer specific questions about the images. In this work, we focus on the alignment stage and identify a significant gap between the embeddings for image and text pairs when VLMs are trained with next-token prediction loss. To address this issue, we employ a contrastive training strategy similar to that used by Radford et al. [39] along with next token prediction based training. Our findings indicate that this joint pretraining method enhances VLM performance by approximately 2% across various multimodal evaluations without any additional compute or training data. To assess the robustness and generalizability of joint training, we experimented with multiple large language models and observed similar performance improvements. Furthermore, we explore the importance of prompts in contrastive training with various LLM options. We also provide a detailed analysis of the type of vision encoder, projection layer, and LLM to use with the proposed joint training approach.

Keywords: Vision Language Models, Multimodal Alignment, Contrastive Training.

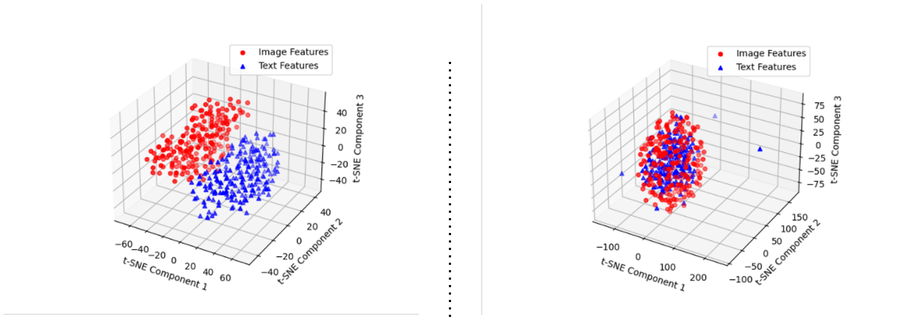
1 Introduction

Large Language Models (LLMs) have achieved state-of-the-art zero-shot performance on a variety of natural language tasks, showcasing their impressive generalization capabilities [13,35,45]. Recently, there has been a growing effort to integrate image modalities with LLMs [1,10,18,22,26,30,33,44], enabling these models to perform tasks such as image classification, optical character recognition, object detection, and more in a zero-shot manner—mirroring their proficiency in solving various NLP tasks without task-specific training. The primary challenge in

[*] Equal contribution

unifying vision and language models for collaborative inference lies in effectively connecting LLMs with vision foundation models. These models are independently pretrained on their respective modalities—text for LLMs and images for vision models.

Liu et al. [30] has recently proposed an effective algorithm to align vision and language models. Their approach involves a two-step process: a pretraining stage and an instruction tuning stage. In the pretraining stage, a projection module is trained to map image embeddings into the language model's space using an image-text pair dataset. During the instruction tuning stage, the model is fine-tuned to answer specific questions about the images, thus enhancing its ability to handle multi-modal inputs. There has been various works studying the impact of using different datasets [33], projection modules such as C-abstractor [7] and LLMs in order to improve the zero-shot multi-modal performance [26].



(a) Text and image embeddings after next token prediction based pretraining.

(b) Text and image embeddings after joint pretraining.

Fig. 1: Figure 1a represents t-SNE visualization [32] of image and text embeddings being unaligned after the next token prediction based pretraining as done by [26, 30]. Figure 1b represents t-SNE visualization of image and text embeddings being aligned after doing joint, contrastive and next token prediction based pretraining.

In this work, we identify that previously proposed pretraining strategies primarily focus on minimizing next token prediction loss, which often results in misaligned vision and language embeddings, as illustrated in Figure 1. Notably, Figure 1a clearly demonstrates that even after pretraining, the embeddings of vision and text remain disparate, contrary to the intended goal of aligning these modalities. This discrepancy underscores the necessity to explore methodologies that not only performs this generative pretraining but also explicitly align vision and text embeddings.

To address the aforementioned discrepancy, we employ an efficient contrastive training strategy inspired by similar approaches in [5, 39]. This strategy aims to minimize the distance between matching image-text pairs while maximizing the separation between dissimilar pairs. We argue that reducing the gap between image and text embeddings could help enable the language model to leverage its extensive pretraining for the vision modality as well.

In order to ensure the model's capability in coherent question answering, we concurrently minimize two losses: contrastive loss and next token prediction loss. As depicted in Figure 1b, joint training objective effectively enhances the model's ability to learn aligned embeddings, where image and text embeddings exhibit closer proximity.

Our findings show that the proposed joint pretraining enhances VLM performance by approximately 2% across various multimodal evaluations without requiring additional computational resources or training data. We experimented with different LLMs to validate the robustness of joint training, observing similar performance boosts over traditional next token prediction loss based pretraining. Additionally, we conducted a detailed analysis on the type of vision encoder, projection layer, and LLM to use for achieving superior results with this joint training approach.

The rest of the paper is organized as follows: Section 2 provides background on LLMs and similarity training. Section 3 details our approach, focusing on the joint training strategy. Section 4 presents the main experimental results. Finally, section 5 summarizes and concludes the paper.

2 Background

2.1 Large language models

Brown et al. [4] showcased the impressive capabilities of GPT-3, a large language model with 175 billion parameters, capable of achieving zero-shot performance comparable to state-of-the-art fine-tuned models. Following this milestone, several other language models have been developed, including PaLM [12], Llama [46], and Falcon [38], among others. These models have consistently demonstrated outstanding zero-shot performance, often matching the performance of fine-tuned models across various datasets.

Ouyang et al. [37] found that although language models can achieve remarkable zero-shot performance on specific benchmarks, they often struggle to accurately understand and follow human instructions when prompted. To address this issue, they proposed a reinforcement learning framework with fine-tuning guided by human feedback, leading to the development of the InstructGPT model. While current LLMs perform exceptionally well on various text-based benchmarks [4], they still struggle to process different modalities, such as audio and images.

2.2 Vision language models

Despite the success of large-scale pretraining and fine-tuning in creating general-purpose language models, developing versatile vision-language models remains challenging due to the independent training of these models. The recently introduced BLIP-2 model [25] has demonstrated ways to align these vision encoder and LLM by introducing Query transformer (Q-Former) component, which helps to map from image embedding space to language model space.

Building on the advancements of BLIP-2, the InstructBLIP work performs comprehensive examination of fine-tuning vision-language models. By training on 13 different datasets, InstructBLIP achieves state-of-the-art results in various vision-language tasks, outperforming both BLIP-2 and the larger Flamingo models [2]. Both BLIP-2 and InstructBLIP uses Vicuna [11] as LLM backbone. The amount of data used to pretrain and instruction-tune InstructBLIP models was approximately 130 million samples. Therefore, a more efficient fine-tuning strategy was needed to align these models with fewer samples.

To address this issue, LLaVa [30] introduces a data distillation mechanism from GPT-4, resulting in a more robust and diverse instruction-tuning dataset that uses significantly fewer samples compared to the InstructBLIP instruction-tuning dataset. This approach achieves state-of-the-art performance on various multimodal benchmarks. Recent work by Lin et al. [26] highlighted that pre-training and instruction-tuning models on both image and text data can degrade performance on text-only tasks. To address this issue, they proposed interleaved image and text pretraining, which enables LLMs to become multimodal while retaining the performance of the original LLM on text-based tasks.

In most of the previous work data plays a key role in improving performance, in this work we first identify that the pretraining strategies used in most previous works lead to unaligned image and text representations as demonstrated in Figure 1. To address this, we adopt a contrastive learning training strategy that enables the model to explicitly learn aligned image and text embeddings.

2.3 Contrastive learning

Contrastive learning has emerged as a powerful approach, enabling models to learn representations by contrasting similar and dissimilar pairs of data points. A notable example is Contrastive Language-Image Pre-Training (CLIP) [40], which leverages large-scale pretraining to align textual and visual representations. CLIP has demonstrated significant improvements in zero-shot learning, setting new benchmarks in various vision-language tasks.

In the realm of self-supervised learning for visual representations, Distillation of Knowledge with No Labels (DINO) [6] employs a teacher-student framework where the student predicts the teacher's output, effectively learning discriminative features without labeled data. Similarly, Simple Framework for Contrastive Learning of Visual Representations (SimCLR) [9] emphasizes data augmentation and introduces a batch-based contrastive learning approach.

In Natural Language Processing (NLP), several works have successfully integrated contrastive learning with LLMs to enhance representation learning and downstream task performance. One notable example is Deep Contrastive Learning for Unsupervised Textual Representations (DeCLUTR) [17], which leverages contrastive learning to train LLMs on large corpora without supervision. Another significant work is Simple Contrastive Learning of Sentence Embeddings (SimCSE) [16], which applies a simple yet effective contrastive learning strategy by using dropout as noise to create augmented views of the same sentence. Additionally, PromptBERT [21] enhances BERT sentence embeddings by utilizing prompts, improving the quality of the embeddings for various tasks.

In a recent study, Jiang et al. [20] utilized contrastive training to address hallucinations in VLMs. Their approach aims to minimize the disparity between non-hallucinatory and hallucinatory text, while also bridging the domain gap between image and text modalities. Their findings highlight the effectiveness of contrastive learning in specifically mitigating hallucinations.

In contrast, our work specifically focuses on reducing the dissimilarity between text and image embeddings to address concerns shown in Figure 1. We show that using this joint training not only produces aligned image-text embeddings but at the same time improves performance of VLM by approximately 2% on different multi-modal benchmarks. We also provide insights into how different projection layers, contrastive training strategies, LLMs and projection layer impacts the performance of this joint training strategy.

3 Approach

Following the LlaVA [30] framework, our approach also utilizes two stages: 1) vision-language alignment and 2) instruction tuning. In this work, we focus on the vision-language alignment stage to address the issue of unaligned image-text embeddings, as illustrated in Figure 1a.

Section 3.1 introduces the model architecture, which incorporates contrastive loss training alongside next-token prediction loss to produce aligned embeddings while preserving the generative capabilities of the LLM. Sections 3.2 and 3.3 detail the explicit loss terms minimized during pretraining to achieve aligned embeddings, as depicted in Figure 1b. In Section 3.4, we design a prompt-based approach for extracting text and image embeddings to enhance contrastive training.

3.1 Model architecture

Figure 2 depicts the model architecture used during the pretraining stage. The architecture demonstrates that contrastive training can be seamlessly integrated without modifying the primary model structure. This allows for joint training using both next-token prediction and contrastive losses. We employ projection weights w_i and freeze all other parameters to align image embeddings with text embeddings, effectively mapping these image embeddings into a space that the language model (LM) can interpret.

This architecture enables the efficient incorporation of image modality into LLMs while maintaining the integrity of the model's existing structure. The following subsection provide further details on contrastive and next-token prediction training.

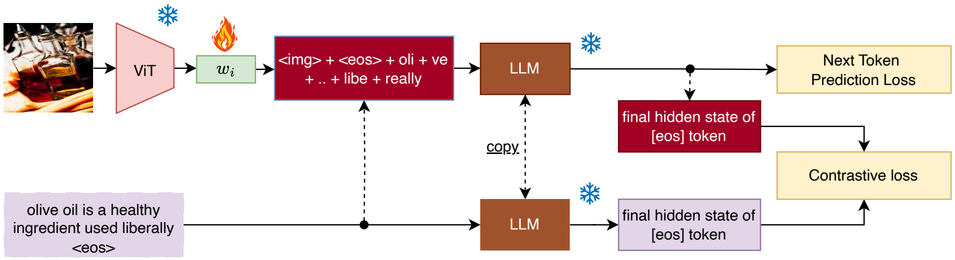

ViT
w_i
<img> + <eos> + oli + ve + .. + libe + really
LLM
Next Token Prediction Loss
final hidden state of [eos] token
copy
Contrastive loss
olive oil is a healthy ingredient used liberally <eos>
LLM
final hidden state of [eos] token

Fig. 2: Detailed architecture of the proposed approach. This figure illustrated pretraining for adding image modality to the language model. We follow a joint training scheme that combines contrastive and generative learning approaches

3.2 Contrastive training

To address the significant differences in image text embeddings, we introduce an additional training scheme that employs contrastive training as demonstrated in Figure 2. The loss defined below aims to align the representations of different modalities more effectively, thereby facilitating improved multimodal understanding:

$$\begin{aligned}\mathcal{L}_c = \frac{1}{2N}\sum_{i=1}^{N}\Bigg[& -\log\frac{\exp(\text{sim}(f(g(x_i)), f(y_i))/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(f(g(x_i)), f(y_j))/\tau)} \\ & -\log\frac{\exp(\text{sim}(f(y_i), f(g(x_i)))/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(f(y_i), f(g(x_j)))/\tau)}\Bigg]\end{aligned} \tag{1}$$

Here, x and y denote the image and text inputs, respectively. The function $f(.)$ refers to the embeddings derived from the LLM's final hidden state at the [EOS] token. As shown in Figure 2 for the image, we first generate image embeddings using a ViT model followed by a projection layer w_i. We then append the [EOS] token to these image embeddings. The function $g(.)$ refers to the embedding derived from the vision modules. The similarity between these embeddings is measured using the function $\text{sim}(\cdot)$, with cosine similarity employed for all samples in a batch of N image-text pairs.

The first term (text-to-image) computes the log probability of the similarity between an image x_i and its corresponding text y_i over the sum of similarities between x_i and all potential texts in the batch. This ensures that the embedding of the image is closely aligned with the correct text embedding while remaining distinct from unrelated text embeddings. The second term (image-to-text) performs a similar computation in reverse, determining the log probability of the similarity between the text y_i and its corresponding image x_i over the sum of similarities between y_i and all other images in the batch. This approach reinforces the alignment between image and text embeddings.

The constant τ is a learnable temperature parameter that controls the sharpness of the similarity distribution. By summing these negative log probabilities and averaging them across the entire batch, the loss function ensures that for each image-text pair in the batch, the correct pairs are significantly more similar than the incorrect pairs.

3.3 Next token prediction training

Similar to most VLMs [30, 47], we incorporate a generative objective based on next-token prediction. In this section, we detail the integration of the next-token prediction loss, specifically implemented as the negative log-likelihood loss, a widely utilized approach in sequence modeling tasks. Within the context of VLM training, this task can be considered as the generative loss $\mathcal{L}_g$, formulated as the negative log-likelihood of all samples in a batch consisting of N image-text pairs:

$$\mathcal{L}_g = -\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T_i} \log P(h_t^i | h_{1:t-1}^i, \mathbf{X}^i) \tag{2}$$

In this equation, h_t^i represents the target token at time step t for the i-th image-text pair, $h_{1:t-1}^i$ denotes the preceding tokens in the sequence for the i-th pair, and $\mathbf{X}^i$ corresponds to the set of image embeddings obtained from the vision encoder followed by projection layer for the i-th image-text pair. The primary objective is to maximize the likelihood of predicting the correct next token for all image-text pairs within the batch. This loss function effectively measures the conditional probability of each word in the caption, given the image features and the preceding words.

3.4 Prompt-based embedding

We further explore the use of prompts in learning embedding representations, following the methodology outlined in PromptBERT [21]. To represent sentences and images with prompts, we utilize a template in the following format: `"[X] means: [EOS]"`, where `[X]` is a placeholder for sentences or images. Based on the best-performing prompt from Jiang et al [21], we construct the following prompts:

- This sentence: {<text>} means: `[EOS]`
- This {photo/image/picture}: {<image>} means: `[EOS]`

Given a {<text>} or {<image>}, we map them using the template. For images, we randomly select one of the three keywords (photo, image, picture). We then feed the prompted input to the language model to generate both sentence and image representations where we employ them during the contrastive training.

4 Experiments

4.1 Implementation details

Our proposed model adheres to LLaVA's implementation for direct comparison purposes [30]. We utilize a two-layer Multi-Layer Perceptron (MLP) to connect vision embeddings to the LLM. The vision encoder is based on Open AI CLIP-L [39] with an image resolution of 336. Alongside Vicuna-7B for direct comparison, we also conduct experiments on various other LLMs, such as Mistral-7B and Llama-8B, to demonstrate that contrastive training can be effective independently of the choice of language model.

We follow 2-stage training as in LLaVA [30]. For pretraining, we use joint loss of $\mathcal{L}_g + \mathcal{L}_c$ with equal weights denoted as w/ cont in all the results below. We pre-extract text embeddings to save computation during the first stage. For intruction tuning stage, we utilize next-token prediction loss to train the models. For both stages, we utilize the same datasets that are constructed in [30].

Stage-1 comprises a dataset of 558K text-image pairs, drawn from a combination of the LAION [41], CC [42], and SBU [36] datasets. For this stage, we employed a batch size of 256 with a learning rate of 2e-3 for 1 epoch. In Stage-2, we fine-tuned the model using a dataset that includes COCO [27], GQA [19], OCR-VQA [34], Text-VQA [43], and Visual Genome [23], with a batch size of 32 and a learning rate of 2e-5 for 1 epoch. To evaluate the robustness of our methods, we tested performance on several datasets, including MMBench (MMB) [31], MMB^{CN} [31], MME [15], Seed-Bench [24], MM-VET [49], and MMMU [50].

4.2 Comparison with state of the art

Table 1: Comparison with State-of-the-Art. Since the introduction of LLaVA [30], several works have been introduced; however, these improvements are largely attributed to the use of larger datasets and more powerful language models [33]. We make direct comparison with LLaVA-7B framework due to its efficient use of dataset and model size.

Model Name	LLM	PT	IT	MMB	MMB^{CN}	MME	Seed	MM-VET	MMMU
BLIP-2 [25]	Vicuna-7B	129M	-	-	-	1293	46.4	22.4	-
InstructBLIP [14]	Vicuna-7B	129M	1.2M	36.0	23.7	-	53.4	26.2	-
Qwen-VL-Chat-7B [3]	Qwen-7B	1.4B	50M	60.6	56.7	1487	58.2	-	32.9
ShareGPT4V-7B [8]	Vicuna-7B	50M	1M	68.8	62.2	1567	69.7	-	37.6
mPLUG-Owl2 [48]	Llama-2-7B	400M	1.2M	64.5	-	1450	60.8	36.2	-
VILA-7B [26]	Llama-2-7B	50M	1M	68.9	61.7	1533	61.1	34.9	-
LLaVA-NeXT-7B [29]	Llama-3-8B	0.6M	0.8M	67.4	-	1519	70.2	43.9	35.8
LLaVA-7B [28]	Vicuna-7B	0.6M	0.7M	64.8	58.3	1510	60.4	31.0	35.1
LLaVA-7B (ours)	Vicuna-7B	0.6M	0.7M	67.2 (↑2.4)	60.4 (↑2.1)	1446 (↓64)	66.9 (↑6.5)	31.4 (↑0.4)	38.6 (↑3.5)

Table 2: Contrastive & prompt comparison for various LLM backbones: These experiments demonstrate that the proposed method generalizes well across various LLM choices, leading to direct model improvements without the need to alter the model architecture or datasets.

LLM Backbone	MMB	MMB^{CN}	MME	Seed	MM-VET	MMMU
Vicuna-7B	64.8	58.3	1510	60.4	31.0	35.0
Vicuna-7B w/ cont	67.2	60.4	1473	66.8	30.0	36.3
Vicuna-7B w/ cont + prompt	67.3	58.5	1446	66.7	31.4	38.6
Mistral-7B	65.4	57.3	1350	66.5	32.4	36.5
Mistral-7B w/ cont	67.2	60.4	1314	64.3	30.1	36.3
Mistral-7B w/ cont + prompt	67.8	60.5	1473	66.7	33.3	38.6
Llama3-8B	73.5	67.7	**1525**	69.7	33.4	39.7
Llama3-8B w/ cont	**74.2**	**68.7**	1487	**69.8**	**36.2**	**41.1**
Llama3-8B w/ cont + prompt	71.1	66.7	1493	68.4	32.1	40.5

We conducted a comprehensive comparison of our model against several state-of-the-art models, as presented in Table 1. It is challenging to make direct comparisons because each model employs different combinations of data, image encoders, and language models. However, our method, LLaVA-7B (ours), can be directly compared to the baseline LLaVA-7B framework since both utilize the same data and architecture, differing only in the training methodology we propose. Compared to other methods, LLaVA-7B performs exceptionally well considering its training data size and efficiency, which is why we use it as our baseline.

As illustrated in Table 1, our model demonstrates a significant boost in most benchmarks compared to LLaVA-7B while maintaining almost the same training time & computation as that of LLaVA-7B. Remarkably, the proposed method performs comparably to LLaVA-NeXT-7B, despite not incorporating larger scale image resolutions, more advanced language model backbones and superior datasets.

4.3 The impact of contrastive pretraining and prompts

Our experiments reveal that incorporating contrastive pretraining (w/ cont) significantly boosts the performance of VLMs. As shown in Table 2, the use of contrastive pretraining across different LLM backbones—Vicuna-7B, Mistral-7B, and Llama3-8B—results in notable performance improvements. We achieve the best results with contrastive pretraining on the Llama3-8B model.

When we enable prompts for embedding extraction (w/ cont + prompt) as described in section 3.4, we observe the most improvements for the Vicuna and

Mistral backbones. However, in Llama3-8B, the use of prompts does not result in a performance boost. This might be due to the fact that the constructed prompt may not be well-suited for that specific backbone. We believe that an optimized prompt for each language model is necessary and can have a substantial impact on performance; we leave this for future work.

4.4 The impact of projection layer

Table 3: Comparison of projection layers: This analysis indicates that the projection layer does not significantly impact overall performance in next token prediction based pretraining (LLaVA-7B data used for all experiments). For the proposed method, there is a notable performance decrease when switching from MLP to average pooling.

Method	MMB	MMB^{CN}	MME	Seed	MM-VET	MMMU
MLP	64.8	58.3	**1510**	60.4	31.0	35.0
C-abstractor [7]	64.3	57.9	1429	62.4	29.2	**37.3**
VILA [26]	64.6	57.9	1322	61.9	29.2	33.8
Avg pooling [33]	64.0	57.0	1380	61.7	28.3	36.0
MLP w/ cont	**67.2**	**60.4**	1473	**66.8**	30.0	36.3
Avg pooling [33] w/ cont	65.9	58.9	1401	61.8	**31.2**	35.8

Table 3 compares the performance of various projection layers. Our analysis indicates that while the choice of projection layer can influence model performance, MLP projections generally maintain high performance levels across the board. One of the main advantages of using projection layers other than MLP is their ability to reduce the number of image embeddings to a fixed number or by a factor, thus optimizing memory usage in the language model and improving inference time. This is particularly important in video processing [30] where multiple images serve as inputs.

We observe that switching to average pooling results in a performance decrease in the proposed method. Specifically, we experiment with an average pooling projection layer, which applies 8×8 average pooling on the output of the ViT image encoder, followed by a linear projection. Our experiments show that adding contrastive training still leads to improvements across various benchmarks; however, the improvement achieved with the MLP layer is more significant. Therefore, our experiments underscore that an MLP projection layer should be favored in most scenarios to maintain optimal performance.

4.5 The impact of vision encoder

Table 4 demonstrates that the choice of ViT encoder significantly impacts the performance of contrastive-trained models. Larger ViTs, such as the CLIP ViT-

Table 4: Comparison of vision encoders: The choice of Vision Transformer (ViT) plays a crucial role in enhancing LLM performance. Larger ViT models generally result in improved performance, and the impact of image resolution is also significant. We use our contrastive training in this ablation study.

Method	MMB	MMB^{CN}	MME	Seed	MM-VET	MMMU
CLIP ViT-B/32	55.2	46.1	1235	56.5	23.1	35.4
CLIP ViT-L/14	62.9	56.2	1360	65.0	28.2	36.0
CLIP ViT-L/14-336	**67.2**	**60.4**	**1473**	**66.8**	**30.0**	**36.3**

L/14, consistently outperform those using smaller encoders like the CLIP ViT-B/32. This observation aligns with findings by McKinzie et al. [33] for next-token prediction-based pretraining. Furthermore, higher image resolutions, as implemented with CLIP ViT-L/14-336, enhance performance metrics. This improvement is intuitive, as models can extract more granular information from high-resolution images.

5 Conclusion

In this study, we have identified that next-token prediction-based pretraining often results in unaligned image and text embeddings. To address this challenge, we introduced a contrastive learning strategy alongside next token prediction loss during the pretraining stage, effectively aligning image-text embeddings. Our experiments demonstrated notable 2% improvements across various multimodal evaluation benchmarks. Additionally, through rigorous ablation studies, we systematically explored the impact of key factors such as the projection layer, LLMs, and vision encoders on our contrastive training strategy. We found that the MLP projection layer consistently outperforms other types of projection layers, and stronger LLM backbones significantly enhance performance. Moreover, using large vision encoders that support higher resolution images proved beneficial for improving VLM performance. We anticipate that our findings will stimulate further research aimed at enhancing the overall performance and versatility of VLMs in multimodal applications.

References

1. Ak, K.E., Lee, G.G., Xu, Y., Shen, M.: Leveraging efficient training and feature fusion in transformers for multimodal classification. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 1420–1424. IEEE (2023) 1
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022) 4
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023) 8
4. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T.J., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. ArXiv **abs/2005.14165** (2020), https://api.semanticscholar.org/CorpusID:218971783 3
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021) 2
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021) 4
7. Cha, J., Kang, W., Mun, J., Roh, B.: Honeybee: Locality-enhanced projector for multimodal llm. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13817–13827 (2024) 2, 10
8. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793 (2023) 8
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020) 4
10. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198 (2024) 1
11. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), https://lmsys.org/blog/2023-03-30-vicuna/ 4
12. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N.M., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B.C., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., García, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A.M., Pillai, T.S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R.,

Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Díaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K.S., Eck, D., Dean, J., Petrov, S., Fiedel, N.: Palm: Scaling language modeling with pathways. ArXiv **abs/2204.02311** (2022), https://api.semanticscholar.org/CorpusID:247951931 3

13. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. Journal of Machine Learning Research **25**(70), 1–53 (2024) 1
14. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023) 8
15. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models (2024), https://arxiv.org/abs/2306.13394 8
16. Gao, T., Yao, X., Chen, D.: Simcse: Simple contrastive learning of sentence embeddings. arXiv preprint arXiv:2104.08821 (2021) 4
17. Giorgi, J., Nitski, O., Wang, B., Bader, G.: Declutr: Deep contrastive learning for unsupervised textual representations. arXiv preprint arXiv:2006.03659 (2020) 4
18. He, M., Liu, Y., Wu, B., Yuan, J., Wang, Y., Huang, T., Zhao, B.: Efficient multimodal learning from data-centric perspective. arXiv preprint arXiv:2402.11530 (2024) 1
19. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019) 8
20. Jiang, C., Xu, H., Dong, M., Chen, J., Ye, W., Yan, M., Ye, Q., Zhang, J., Huang, F., Zhang, S.: Hallucination augmented contrastive learning for multimodal large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27036–27046 (2024) 5
21. Jiang, T., Jiao, J., Huang, S., Zhang, Z., Wang, D., Zhuang, F., Wei, F., Huang, H., Deng, D., Zhang, Q.: Promptbert: Improving bert sentence embeddings with prompts. arXiv preprint arXiv:2201.04337 (2022) 5, 7
22. Kim, T., Ahn, P., Kim, S., Lee, S., Marsden, M., Sala, A., Kim, S.H., Han, B., Lee, K.M., Lee, H., et al.: Nice: Cvpr 2023 challenge on zero-shot image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop: New Frontiers for Zero-Shot Image Captioning Evaluation. pp. 7356–7365 (2024) 1
23. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**, 32–73 (2017) 8
24. Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., Shan, Y.: Seed-bench: Benchmarking multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13299–13308 (2024) 8
25. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022) 3, 8
26. Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., Han, S.: Vila: On pre-training for visual language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26689–26699 (2024) 1, 2, 4, 8, 10

27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 740–755. Springer International Publishing, Cham (2014) 8
28. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26296–26306 (2024) 8
29. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/ 8
30. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024) 1, 2, 4, 5, 7, 8, 10
31. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? arXiv preprint arXiv:2307.06281 (2023) 8
32. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(86), 2579–2605 (2008), http://jmlr.org/papers/v9/vandermaaten08a.html 2
33. McKinzie, B., Gan, Z., Fauconnier, J.P., Dodge, S., Zhang, B., Dufter, P., Shah, D., Du, X., Peng, F., Weers, F., et al.: Mm1: Methods, analysis & insights from multimodal llm pre-training. arXiv preprint arXiv:2403.09611 (2024) 1, 2, 8, 10, 11
34. Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A.: Ocr-vqa: Visual question answering by reading text in images. In: 2019 international conference on document analysis and recognition (ICDAR). pp. 947–952. IEEE (2019) 8
35. Mohta, J., Ak, K., Xu, Y., Shen, M.: Are large language models good annotators? In: Advances in Neural Information Processing Systems Workshop: I Can't Believe It's Not Better- Failure Modes in the Age of Foundation Models 1
36. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. In: Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 24. Curran Associates, Inc. (2011), https://proceedings.neurips.cc/paper_files/paper/2011/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf 8
37. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L.E., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., Lowe, R.J.: Training language models to follow instructions with human feedback. ArXiv **abs/2203.02155** (2022), https://api.semanticscholar.org/CorpusID:246426909 3
38. Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., Launay, J.: The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. arXiv preprint arXiv:2306.01116 (2023), https://arxiv.org/abs/2306.01116 3
39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 1, 2, 8
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 4

41. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs (2021), https://arxiv.org/abs/2111.02114 8
42. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Gurevych, I., Miyao, Y. (eds.) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-1238, https://aclanthology.org/P18-1238 8
43. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019) 8
44. Sukhbaatar, S., Golovneva, O., Sharma, V., Xu, H., Lin, X.V., Rozière, B., Kahn, J., Li, D., Yih, W.t., Weston, J., et al.: Branch-train-mix: Mixing expert llms into a mixture-of-experts llm. arXiv preprint arXiv:2403.07816 (2024) 1
45. Tay, Y., Dehghani, M., Tran, V.Q., Garcia, X., Wei, J., Wang, X., Chung, H.W., Shakeri, S., Bahri, D., Schuster, T., et al.: Ul2: Unifying language learning paradigms. arXiv preprint arXiv:2205.05131 (2022) 1
46. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. ArXiv **abs/2302.13971** (2023), https://api.semanticscholar.org/CorpusID:257219404 3
47. Tsimpoukelli, M., Menick, J.L., Cabi, S., Eslami, S., Vinyals, O., Hill, F.: Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems **34**, 200–212 (2021) 7
48. Ye, Q., Xu, H., Ye, J., Yan, M., Hu, A., Liu, H., Qian, Q., Zhang, J., Huang, F., Zhou, J.: mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration (2023), https://arxiv.org/abs/2311.04257 8
49. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities (2023), https://arxiv.org/abs/2308.02490 8
50. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9556–9567 (2024) 8