Closing the Gap of Latent Spaces for Image-Text Arithmetic

Minh Vu
21minh.vb@vinuni.edu.vn
CS, VinUniversity
Vietnam

Balashova Ekaterina
21ekaterina.b@vinuni.edu.vn
CS, VinUniversity
Vietnam

Tran Anh Vu
21vu.ta@vinuni.edu.vn
CS, VinUniversity
Vietnam

1 Introduction

The rapid advancement of artificial intelligence (AI) has led to remarkable applications in fields like chatbots, autonomous vehicles, and recommendation systems. A key factor behind this success is enabling machines to understand and interpret essential data such as images, text, and voice for tasks like decision-making and content generation. At the core of this process are embeddings, which is numerical representations of data that allow machines to interpret and process complex information. The quality of these embeddings directly impacts the learning efficiency, with more accurate embeddings leading to better model performance.

In recent years, there has been growing interest in combining embeddings from different data types, such as images and text, in order to provide more diverse and comprehensive information to AI systems. This has been a significant step toward improving AI's learning performance. Vision-Language Models (VLMs) are a prominent example of this, where visual and textual data are integrated to enhance the machine's understanding of both modalities simultaneously. One of the most notable models in this space is **CLIP (Contrastive Language-Image Pre-training)** [14], which generates separate embeddings for images and text. CLIP has made significant progress in aligning image and text representations, but these two spaces are not fully shared, meaning that the relationship between the textual and visual representations is not entirely "mixed" together. This gap can be attributed to the inherent differences in the nature of the data and the initial relationships between instances within each modality. Therefore, in this project, we aim to **explore the modality gap between text and image representations within CLIP's embedding space.** By applying a novel translation method, we seek to create better-shared embeddings that align both the visual and textual data, ultimately improving the performance of multimodal AI systems. Additionally, this project supports the development of visual arithmetic capabilities, enabling AI models to perform tasks that involve reasoning with visual and textual information simultaneously. This could lead to more advanced applications in areas like visual question answering, image captioning, and even complex reasoning tasks where both modalities need to be considered together.

2 Literature Review

Having been mentioned as "distributed representations" for words by Bengio et al. [1] in the early 2000s, the modern idea of text embeddings, especially embeddings for words, became prominent with the introduction of Word2Vec by Tomas Mikolov [12] and his team at Google in 2013. By extracting the weight as the text representation, Word2Vec can capture both syntactic and semantic relationships after the neural network training process to predict the word given context or vice versa. Working on the corpus scale, GloVe [13], on the other hand, refines the process by using matrix factorization to incorporate global word co-occurrence statistics, providing a richer word representation. While these models struggled with capturing more nuanced semantic relationships, such as analogies and contextual variations, due to their static and inflexible embedding methods, they still laid a strong foundation for the development of a robust framework that captures semantic and syntactic relationships in text.

One approach that enhance the model capability of semantic meaning capturing is providing more diverse information from cross-modal data, in which people incorporate different data types leading to the development of multimodal embedding spaces. These spaces allow for the integration of multiple modalities, such as images and texts, enabling models to perform tasks like image-text and text-image retrieval [6] , or zero-shot classification [3], where cross-modal understanding becomes essential. Among those models, CLIP, introduced in 2021, has become an outstanding baseline in the field. Leveraging a dual-encoder approach, where both images and text are embedded into a shared space to learn their relationships through contrastive pre-training, the framework creates a dataset classifier using textual labels, then performs zero-shot prediction by associating new images with their most relevant textual description. From this, CLIP can then be used as a generalized backbone for several different downstream tasks such as captioning, vision question answering, or even more low-level usage because we can also make use of its pre-trained vision-text encoder. Additionally, re-training or even fine-tuning CLIP with an efficiently large dataset seems too costly for individual projects, so some other methods of low resources are also examined. To address that issue, LiT (Locked-image Tuning), which freezes a powerful pre-trained image encoder and trains only a text encoder via contrastive alignment on image–text pairs, or SigLIP [20], which replaces CLIP's softmax-based contrastive loss with a pairwise sigmoid loss, treating each image–text pair independently, are potential solutions to avoid large batch sizes.

While CLIP is impressive, it suffers from bad compositional reasoning, fine grained understanding of image content, as mentioned in [2] and [9]. In addition to this, we can still find the gap in the alignment between different modality embeddings space, especially through our experiment of visual arithmetic capability. Realizing the same issue, ZeroCap [18] claims an impressive visual arithmetic capability that allows the numerical result produced by visual/ text or multimodal embeddings addition or subtraction to hold desired corresponding visual/text meanings. Specifically, [18] has done it by directly steering the context vectors (cache) Ci closer to the

CLIP's visual embedding at inference time instead of training the language model itself. This technique makes the image description output more case-specific, more diverse, creative, and closer to the expectation of arithmetic outcomes, compared to the self-supervised learning approaches. However, essentially this approach makes no change to CLIP's embedding space - CLIP's embedding space and the language model's embedding space are disconnected. Moreover, it is limited in assessing cross-modality embedding space and scaling to improvements in downstream tasks. With the idea of enhancing the textual embedding space of CLIP, JinaCLIP [8] improves the semantic meanings of those embeddings specifically in text-only tasks by additional training on synthetic text datasets (pairs of image-long captions generated by other language models (e.g. GPT4v)), This approach, while increasing the VLM's model capability to understand text meaning, has not thoroughly improved the similarity of embeddings representing the same thing in visual or cross-modal spaces.

Inspired by the aforementioned success of ZeroCap, we want to reduce the modality gap directly to ensure downstream applications. To achieve that, instead of improving embedding encoders, performing statistical methods, e.g. translational mapping, on existing representation might be a better choice in terms of computational resources. While the idea of embedding a linear shifting method aiming to make 2 sets of text and visual vectors closer was introduced by [10], this approach was later proved to be inefficient by [3] which also failed to address the issue because of the dataset's lack of image description details. Another reason for the gap between different embedding coming from different modalities pointed out by [15] is: the distinguished structure of data within each modality which makes the linear mapping method ineffective.

Consequently, for that rationale, we aim to develop a non-linear translation mapping (which could be done via a neural network with non-linear activation functions) that can directly align images and text, which enables cross-modal arithmetic. By adjusting the detail level of text data, we can evaluate the corresponding impact on the resulting modality gap.

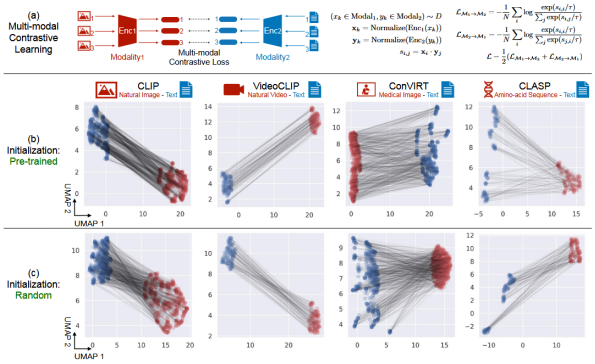

Figure 1: Modality gap in multimodal contrastive representation learning [10]

3 Methodology

Our idea is to train a model that would translate CLIP image embeddings into the corresponding text embeddings in an attempt to close the modality gap.

3.1 Datasets and Preprocessing

3.1.1 Data Description. This project uses 3 datasets:

(1) CC3M - Conceptual Captions [16]. This dataset contains 3,3 million images with accompanying captions that were automatically mined from the internet, creating a large amount of diverse data. It is very popular for image captioning tasks.

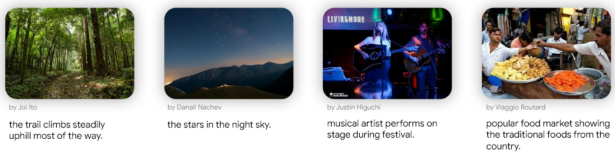

Figure 2: Sample images from the CC3M dataset

(2) COCO - Common Objects In Context [11]. This dataset contains 328,000 images of common objects, each manually labeled with 5 captions. This highly curated dataset is well-known and widely used in image captioning tasks.

Figure 3: Sample image from the COCO dataset

(3) DCI - Densely Captioned Images[19]. This dataset contains 7805 images with very detailed captions. We were able to use only about 700 datapoints from this dataset due to the high inference cost.

From the CC3M dataset, we are using 10,000 images for model training. We are also using the entire DCI dataset for that purpose. From COCO, we are using the 2017 validation split (5,000 images) as our test set.

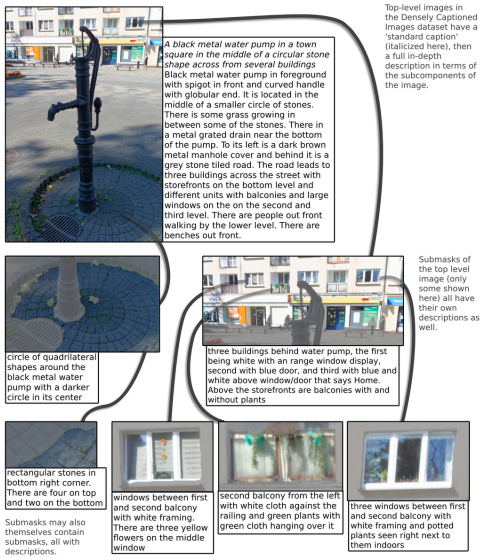

Figure 4: Sample image from the DCI dataset

3.1.2 Preprocessing Method. The data is preprocessed the following way:

(1) For the images in CC3M and COCO datasets, synthetic captions are generated using Gemini 2.0 Flash Lite. The rationale for it is that [2], [15], and [19] showed that due to human subjectivity, VLMs tend to produce more high-quality captions compared to human captioners, especially so in the case of web-crawled captions like in CC3M.
(2) The final images and corresponding captions are used as inputs to CLIP to get their embeddings.

3.2 Model and Benchmark

We are using a simple model architecture to learn to align image embeddings with the text embeddings in the CLIP space. Our main inspiration for this is the evidence that the innate relationship and structure of each modality space is not identical [15], which makes linear translation methods ineffective [10] [4]. We use the simple Mean Squared Error loss, which was shown to be similar to minimizing the cosine distance in [10] Therefore, we opted to investigate the potential of a fairly simple nonlinear transformation. The original CLIP embeddings serve as the baseline to evaluate the changes introduced. The details of the model implementation and the experiments can be found on our GitHub.

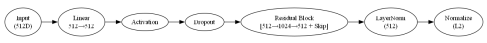

Figure 5: Non-Linear Aligner architecture

3.3 Arithmetics

While we investigate the cross-modal arithmetic ability of VLMs, we come across a prominent work of ZeroCap [18] that claims to enable visual arithmetic. However, deeper scrutiny of this work reveals that ZeroCap is only imitating this ability through an external language model. In other words, it utilizes the innate alignment of CLIP by trying to generate the text that is closest to the queried image. Regardless, this work shows that CLIP's embedding space has potential to enable cross-modal arithmetic given the sound alignment of the space.

We believe that the existence of modality gap critically hinders the cross-modal arithmetic. Since each modality is pushed to a small cone-like region, performing arithmetic using their vectors would just result in an empty, secluded space. Thus, the novelty of our approach is to directly align image embeddings with the text embeddings. More precisely, we use the text embeddings to replace corresponding image embeddings, while learning to map new images into the text space using our model. This is a simple yet effective way to solve this issue.

4 Evaluation and Experimental results

4.1 Modality Gap

To evaluate how well our approach contributes to closing the modality gap, we are using the three key modality gap metrics:

(1) Alignment score — average cosine similarity between matching text and image embeddings, proposed by Goel et al. (2022)[5] and widely used .

$$\text{Alignment} = \frac{1}{N}\sum_{i=1}^{N}\cos(\mathbf{x}_i, \mathbf{y}_i) \tag{1}$$

(2) L2M — L2 distance between modality means. Proposed by Liang et al. (2022) [10] and popular as the simplest metric.

$$\text{L2M} = \left\| \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i - \frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i \right\|_2 \tag{2}$$

(3) Relative Modality Gap (RMG) — the proportion of mean distance between matching pairs to the sum of itself and mean intra-modality distance. Proposed by Schrodi et al. (2024) [15].

$$\text{RMG} = \frac{\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i - \mathbf{y}_i\|_2}{\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i - \mathbf{y}_i\|_2 + \frac{1}{2N(N-1)}\left(\sum_{i\neq j}\|\mathbf{x}_i - \mathbf{x}_j\|_2 + \sum_{i\neq j}\|\mathbf{y}_i - \mathbf{y}_j\|_2\right)} \tag{3}$$

The results are presented in Table 1. As evident, our approach leads to improvement across all metrics. The embedding space also reflects these changes (Figure 6).

Table 1: Evaluation metrics before and after training

Metric	Original CLIP	After Non-Linear Aligner
Alignment Score (↑)	0.3455	0.7607
L2M (↓)	0.7508	0.2070
RMG (↓)	0.5138	0.4032

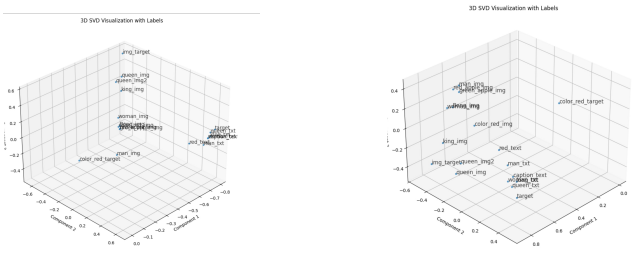

Figure 6: SVD of the embedding space before and after adjustment

The embeddings of texts and images are better mixed after our adjustment.

5 Discussion

5.1 Limitations

One major limitation of our model is our model is highly-dependent on the capability of the text modality's embedding space. We are using the CLIP to embed our data so the performance can be limited by CLIP's performance. We believe we can achieve better results were we to use a model that has better inter-modality alignment from the beginning, like Jina CLIP or Jina CLip v2, which achieves it through improving text embeddings [8], [7], or AlignCLIP, specifically designed to mitigate the modality gap [4]. However, we were unable to gain sufficient and stable access to these model, so default CLIP was used. The usage of this model also introduces another limitation: the maximum length for text inputs is only 77 tokens, which is insufficient for truly dense captions, limiting the quality of text embeddings.

Another limitation is that due to the limited resource availability, we were only able to work with small subsets of the datasets, which inevitably limited the quality of the model.

Finally, our testing was done within CLIP's embedding space, which is affected by information imbalance, while our data mitigate information imbalance of the text data. This makes our model dependent on how detailed text data are during testing. Because of the large scope of this problem with heated discussion within the topic itself, we focused the majority of our time digging into the problem to come up with our own efficient and easy-to-train solution. The lack of resources also urged us to look at different training settings, such as SigLIP or Locked-image Tuning (LiT), and different approach, like statistical intra- and inter-modality shifts or shared architecture like AlignCLIP. Therefore, we did not have sufficient time to evaluate our findings on better base embedding models and on downstream tasks like zero-shot classification and retrieval tasks.

6 Conclusion & Future Work

While our approach does contribute towards closing the modality gap and improve cross-modal arithmetic, there is much room for further investigation and improvement. One interesting direction would be investigating the performance of our modification on zero-shot classification and image-text or text-image retrieval task to see how it compares to the original model. We can also look into using more complex models to introduce non-linearity. Another one is to explore this method on other CLIP-based models, including the use of sigmoid loss by SigLIP [20] or weighted contrastive loss with CWCL [17].

7 Members Contribution

- **Binh Minh:** Team leader. Found and distributed literature, read and presented a lot of it. Participated in ideation. Ran some tests. Developed final model architecture.
- **Kate:** Handled data finding, loading, and preprocessing. Read and presented some literature. Participated in ideation. Ran some tests.
- **Vu:** Read and presented some literature. Participated in ideation. Kept track of team's findings and ensured the project stayed on track. Ran some tests.

References

[1] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A Neural Probabilistic Language Model. In *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich, and V. Tresp (Eds.), Vol. 13. MIT Press. https://proceedings.neurips.cc/paper_files/paper/2000/file/728f206c2a01bf572b5940d7d9a8fa4c-Paper.pdf

[2] 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. 2023. Dense and Aligned Captions (DAC) Promote Compositional Reasoning in VL Models. arXiv:2305.19595 [cs.CV] https://arxiv.org/abs/2305.19595

[3] Sedigheh Eslami and Gerard de Melo. 2024. Mitigate the Gap: Investigating Approaches for Improving Cross-Modal Alignment in CLIP. arXiv:2406.17639 [cs.CV] https://arxiv.org/abs/2406.17639

[4] Sedigheh Eslami and Gerard de Melo. 2025. Mitigate the Gap: Improving Cross-Modal Alignment in CLIP. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://github.com/sarahESL/AlignCLIP Poster.

[5] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. 2022. CyCLIP: Cyclic Contrastive Language-Image Pretraining. arXiv:2205.14459 [cs.CV] https://arxiv.org/abs/2205.14459

[6] Andrej Karpathy and Li Fei-Fei. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. arXiv:1412.2306 [cs.CV] https://arxiv.org/abs/1412.2306

[7] Andreas Koukounas, Georgios Mastrapas, Sedigheh Eslami, Bo Wang, Mohammad Kalim Akram, Michael Günther, Isabelle Mohr, Saba Sturua, Nan Wang, and Han Xiao. 2025. jina-clip-v2: Multilingual Multimodal Embeddings for Text and Images. arXiv:2412.08802 [cs.CL] https://arxiv.org/abs/2412.08802

[8] Andreas Koukounas, Georgios Mastrapas, Michael Günther, Bo Wang, Scott Martens, Isabelle Mohr, Saba Sturua, Mohammad Kalim Akram, Joan Fontanals Martínez, Saahil Ognawala, Susana Guzman, Maximilian Werk, Nan Wang, and Han Xiao. 2024. Jina CLIP: Your CLIP Model Is Also Your Text Retriever. arXiv:2405.20204 [cs.CL] https://arxiv.org/abs/2405.20204

[9] Martha Lewis, Nihal V. Nayak, Peilin Yu, Qinan Yu, Jack Merullo, Stephen H. Bach, and Ellie Pavlick. 2024. Does CLIP Bind Concepts? Probing Compositionality in Large Image Models. arXiv:2212.10537 [cs.CV] https://arxiv.org/abs/2212.10537

[10] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. arXiv:2203.02053 [cs.CL] https://arxiv.org/abs/2203.02053

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312 [cs.CV] https://arxiv.org/abs/1405.0312

[12] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs.CL] https://arxiv.org/abs/1301.3781

[13] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[14] 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. arXiv:2103.00020 [cs.CV] https://arxiv.org/abs/2103.00020

[15] Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. 2025. Two Effects, One Trigger: On the Modality Gap, Object Bias, and Information Imbalance in Contrastive Vision-Language Models. arXiv:2404.07983 [cs.CV] https://arxiv.org/abs/2404.07983

[16] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-Text Dataset for Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2556–2565.

[17] Rakshith Sharma Srinivasa, Jaejin Cho, Chouchang Yang, Yashas Malur Saidutta, Ching-Hua Lee, Yilin Shen, and Hongxia Jin. 2023. CWCL: Cross-Modal Transfer with Continuously Weighted Contrastive Loss. arXiv:2309.14580 [cs.LG] https://arxiv.org/abs/2309.14580

[18] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2022. ZeroCap: Zero-Shot Image-to-Text Generation for Visual-Semantic Arithmetic. arXiv:2111.14447 [cs.CV] https://arxiv.org/abs/2111.14447

[19] Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. 2024. A Picture is Worth More Than 77 Text Tokens: Evaluating CLIP-Style Models on Dense Captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 26700–26709.

[20] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid Loss for Language Image Pre-Training. arXiv:2303.15343 [cs.CV] https://arxiv.org/abs/2303.15343