Self-Aligning VLMs with Focus on Image Modality

Aayush Bajaj [*1,2] **Milosh Devic** [*1,2] **Carlone Scott** [*1,2] **Paul Kelendji** [*1,2]

Abstract

Vision-language multimodal models (VLMs) have recently gained a lot of attention given their ability to tackle complex tasks like visual commonsense reasoning, scene understanding, and spatio-temporal reasoning. This performance of VLMs significantly depends on their 'alignment' abilities: the ability of the models to create precise mappings between the visual and the textual components. Hence, there is growing interest in formulating innovative methods for improving image-text alignments, specifically focusing on improving the images' representation space. Existing methods aim at augmenting the existing texts present in the alignment training data, to make the texts more informative for effective image-text alignment (Doveh et al., 2023a). However, these augmentation methods generally are not grounded to the images, often leading to either (i) insufficient information in the text for the given image, or (ii) overgenerating text which contains extra hallucinated information. To this end, we introduce a novel yet simple method for generating faithful augmentations for captions in image-text alignment training data by back-generating images using the augmented texts and conditioning on the original images, leading to the creation of better grounded image-text alignment data and tackling both ends of the issue.

1. Introduction

In recent times, significant advancements in the fields of generative language (OpenAI, 2023) (Touvron et al., 2023) and vision modelling (Rombach et al., 2022), coupled with the development of Vision & Language (VL) alignment methodologies, have brought us significantly closer to creating models that fully grasp and comprehend the world around them. These models are also capable of acting based on such comprehension. The VL models, serving as the 'eyes' of these innovative systems, have demonstrated robust performance across a variety of tasks, including but not limited to, recognition, detection, segmentation, VQA, and captioning, among others.

The current discourse predominantly emphasizes enhancing the text modality particularly for captioning, yielding remarkable improvements in model performance across various applications (Dunlap, 2023) (Yang, 2023). However, the potential of focusing on the image modality remains relatively underexplored, presenting an opportunity to pioneer a novel approach in the field.

This research proposal introduces an innovative methodology aimed at enriching the image modality within VL models. Our approach centers on generating images from existing captions within the LAION dataset, a subset of the Comprehensive Conceptual Captioning (CC3M) dataset. By leveraging generative techniques to create images that are inherently aligned with their textual descriptions, we aim to produce a more faithful representation of the described scenes (Azizi, 2023). The generated image-text pairs will then be utilized to retrain VL models, hypothesizing that this process will lead to enhanced model comprehension and performance by providing a richer, more accurate visual context.

The motivation behind this approach is twofold. Firstly, by augmenting the image modality, we anticipate overcoming some of the limitations posed by the existing biases towards text modality improvements. Secondly, this strategy seeks to harness the untapped potential of image generation from text to foster a deeper, more nuanced understanding of the visual world by VL models.

Relevant literature underscores the disparity in the development pace between text and image modalities in VL research. For instance, advancements in text generation and understanding have been propelled by significant works in generative language modeling, while parallel progress in synthetic data for VL models (Yu et al., 2023) (Azizi, 2023) has been limited to Visual Recognition and classification

[*]Equal contribution [1]Université de Montréal, Canada [2]Mila - Quebec AI Institute, Canada. Correspondence to: Aayush Bajaj <aayush.bajaj@umontreal.ca>, Milosh Devic <milosh.devic@umontreal.ca>, Carlone Scott <carlone.scott@umontreal.ca>, Paul Kelendji <paul.kelendji@umontreal.ca>.

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

tasks. However, VL alignment techniques (Doveh et al., 2023a) highlight the potential for comprehensive model capabilities. This proposal aims to bridge this gap by specifically targeting the enhancement of the image modality, thereby contributing to the equilibrium in modality advancements.

In conclusion, this research seeks to not only address the current asymmetry in VL model development but also to explore the theoretical and practical implications of a balanced approach to image and text modalities. By focusing on generating and utilizing image-text pairs more faithfully, we aspire to pave the way for VL models that offer a more holistic understanding and interaction with the world, setting a precedent for future research in the domain.

2. Related Work

Research in VLMs has significantly progressed, aiming to enhance the alignment between visual and textual data. Models like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) have pioneered learning robust multimodal representations from large-scale datasets, demonstrating the power of natural language supervision in improving zero-shot learning capabilities. These advancements underscore the importance of precise image-text alignment for tasks such as image captioning and visual question answering.

However, while substantial progress has been made in text-to-image coherence, the reciprocal process—generating images that are semantically aligned with text descriptions—remains less explored. Techniques for text augmentation in image-text datasets (Doveh et al., 2023b) highlight the potential for improving alignment through enriched textual descriptions but do not address the visual modality directly. Conversely, generative models like DALL·E (Ramesh et al., 2021) offer promising avenues for creating visually rich content from textual inputs, suggesting a potential method for enhancing the image modality in VLMs.

Our project aims to bridge this gap by focusing on the generation of contextually relevant images from text descriptions, leveraging the untapped potential of image modality enhancement to advance the state-of-the-art in multimodal alignment. By doing so, we seek to contribute a novel perspective to the ongoing discourse in the field, emphasizing the need for a balanced approach to improving both modalities in VLMs.

3. Methodology

We will present in short the general idea of our methodology for this project (see Figure 1). We will go more into detail for each part in the following sections.

The data collection process involves gathering BLIP captions from an existing image-text dataset and then generating new image-caption pairs using an appropriate open-source model. Subsequently, a pre-trained CLIP model is employed to calculate scores for each pair, and those with low CLIP scores are removed to ensure alignment and capture of the essence of the captions. Following this, the remaining image-caption pairs are fed into the VLM model. The VLM model is iteratively optimized by fine-tuning and adjusting data selection criteria based on evaluation results until satisfactory performance is achieved. Finally, the model is evaluated on a separate validation set, and if successful, it is deployed for generating captions for new images in real-world scenarios.

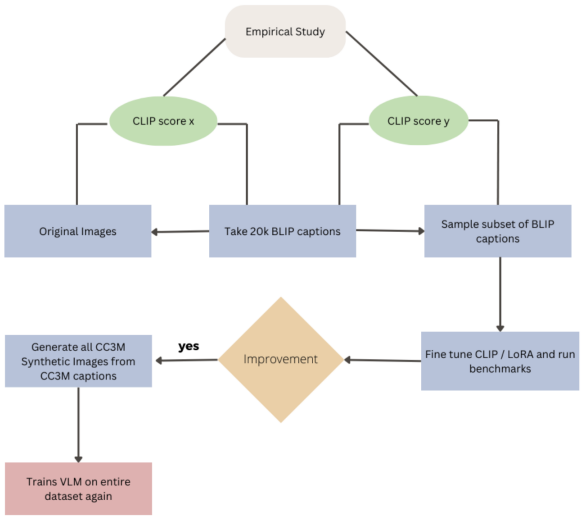

Figure 1. Procedure for VLM Improvement.

4. Synthetic Dataset

As we can see in 3, we have a diagram showing the workflow for creating our synthetic dataset. Initially, we wanted to have about 20,000 image-caption pairs but it was too large to handle by one person at once (OOM errors would come up), therefore, we had to implement a variety of strategies and methods to mitigate this issue.

4.1. Data Collection:

First, we start with the LAION dataset. We randomly sample 4 subsets of 5,000 images from it. These sampled images are then copied to a designated destination directory, with the script ensuring that no duplicates are overwritten. This subset will serve as the basis for our synthetic dataset generation. Following this, a custom function is utilized to extract the BLIP captions from a JSON file containing metadata

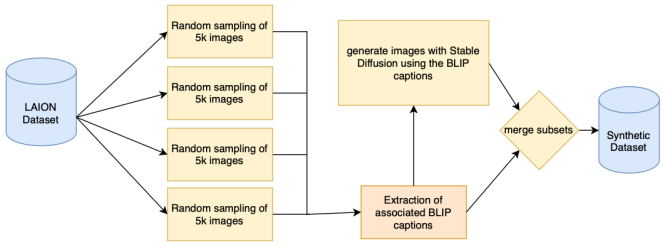

Figure 2. Workflow of the Creation of our Synthetic Dataset

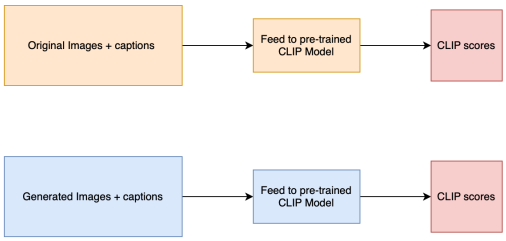

Figure 3. CLIP Scores Calculation Diagram

associated with the images. The function iterates through the metadata, matching image IDs with those of the sampled images, and collects the corresponding captions. Finally, the extracted captions are saved to a new JSON file for further analysis. This method provides a systematic approach to obtain a manageable subset of images along with their associated textual data, facilitating downstream tasks such as image-captioning models or image retrieval systems.

4.2. New Image-Caption Pairs:

For each of the 5,000-image subsets, we utilized Stable Diffusion, a state-of-the-art image generation model. We generate new images based on the captions associated with the original images. To do this step, we needed to implement several different methods to make it work because this is the step where we were getting OOM errors:

- We used "torch.cuda.empty_cache()" as a 'clean-up' command in our code. It frees up memory that's no longer needed, like closing unused programs on a computer to speed things up. By using it, we ensure our deep learning process runs smoothly without memory-related slowdowns.
- We initialized the scheduler and the Stable Diffusion 2 base pipeline with pre-trained components. We used this version of Stable Diffusion because it's for images of size 512x512 whcih is compatible for CLIP and less heavy on the memory than Stable Diffusion 3 (which is 768x768).
- We used torch.float16 as the data type (torch_dtype) for the Stable Diffusion pipeline because it helps reduce memory usage. Float16 (half-precision floating-point format) consumes less memory compared to the default torch.float32 (single-precision floating-point format). While float16 may have lower precision, it's sufficient for our task and helped reduce memory requirements, allowing us to work with larger models and batches.
- The parameter "use_safetensors=True" in the "StableDiffusionPipeline" constructor enables the use of safe tensor operations. Safe tensors optimize memory usage by avoiding unnecessary memory allocations and deallocations during tensor operations. This optimization helps reduce memory fragmentation and overhead, further conserving resources.
- We processed the data in batches rather than individually to minimize memory overhead. Batching allows us to perform computations on multiple inputs simultaneously, maximizing hardware utilization and reducing the overall memory footprint per operation. This is particularly important when working with large models/datasets, as it helps ensure that memory usage remains manageable, even on devices with limited resources like GPUs.
- Finally, we used garbage collection process to reclaim memory occupied by objects that are no longer in use, freeing up resources and reducing memory usage.

Once we had our generated images and extracted captions, we merged these subsets. This merging creates a consolidated pool of images and text, which will form our synthetic dataset.

4.3. CLIP Scores:

Before proceeding further, we aimed to assess our dataset using CLIP scores to gauge the alignment of the new image-caption pairs. Figure 3 illustrates this simple step.

CLIP scores offer a powerful means of assessing the semantic alignment between images and their corresponding captions. Leveraging a pre-trained model that simultaneously learns to understand both textual and visual information, CLIP provides a quantitative measure of how well an image and its associated caption capture similar semantic concepts. By incorporating CLIP scores into our evaluation, we gain valuable insights into the effectiveness of our dataset in capturing meaningful correlations between images and their

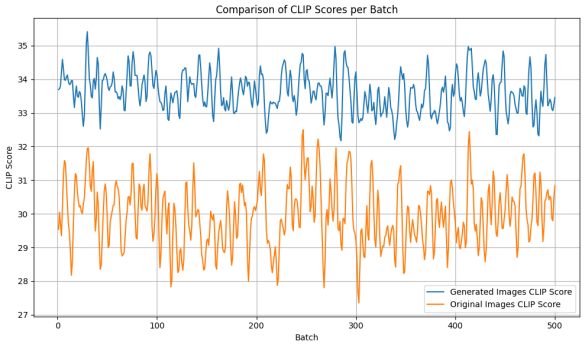

Figure 4. Clip Scores comparison distribution among generated images and original images (subset 1 of 5k images

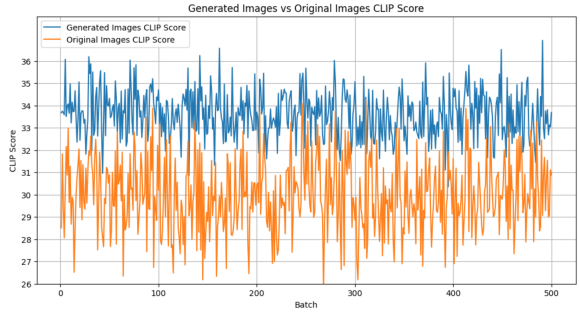

Figure 5. Clip Scores comparison distribution among generated images and original images (subset 2 of 5k images)

descriptive text. This step is crucial in ensuring the quality and coherence of our dataset, laying a solid foundation for downstream tasks such as image captioning and retrieval.

In our preliminary tests involving CLIP score distribution among the image caption pairs from synthetic and original samples (figures 4 and 5), we have concluded that synthetic images are aligned well with their descriptive captions or prompts.

5. Training:

The training component of our project focuses on the fine-tuning of the CLIP model (Radford et al., 2021) using a low-rank adaptation approach (Hu et al., 2021), which has shown promising results in enhancing model performance with minimal computational overhead. Our training methodology is designed to leverage the synthetic image-caption pairs generated in earlier stages of the research, aiming to improve the alignment and comprehension capabilities of the vision-language models. The following steps are then taking for pre-processing:

- Initially, the synthetic dataset, comprising approximately 20,000 image-caption pairs, was split into training, validation, and test sets (see 6). This division was crucial for assessing the model's performance incrementally and ensuring the robustness of the training process. The data then goes through 2 preparation processes (see figure 6):

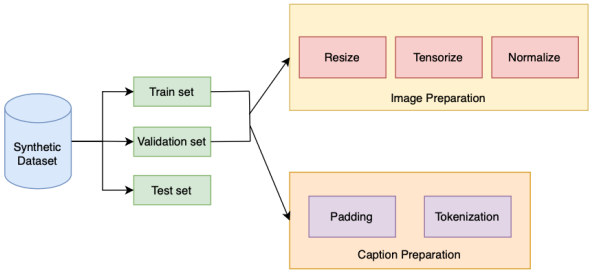

Figure 6. Pre-processing Step

- On the image side, we perform a series of standard pre-processing steps: we resize the images to ensure uniformity, tensorize them to convert into a format suitable for us and normalize the image tensors to have values that aid in faster and more stable training.
- The captions underwent padding to ensure uniform length and were tokenized to transform the textual data into a format that our model could effectively process.

These pre-processing steps were crucial for minimizing potential sources of bias and variance that could affect the training outcomes. Using the prepared dataset, the CLIP model underwent fine-tuning through a technique called Low Rank Adaptation (see figure 7) . This approach selectively updates a small subset of model parameters, thus allowing for efficient training with significantly reduced computational resources. We hypothesized that even with ranks as low as 1, substantial improvements in model performance could be observed, given the targeted nature of the updates.

The model is be evaluated based on the VL-checklist listed in (Zhao et al., 2023) which is a popular method used nowadays to compare vision-language models. This helps us evaluate our method against other techniques that have used textual modality to improve VLM models.

Throughout the training process, we encountered several challenges, primarily related to memory management and computational efficiency. To address these, we implemented strategies such as batch processing and memory optimization techniques, which included using lower precision data formats and optimizing tensor operations. These adjustments were critical in managing the large-scale data and complex model architectures involved in our study.

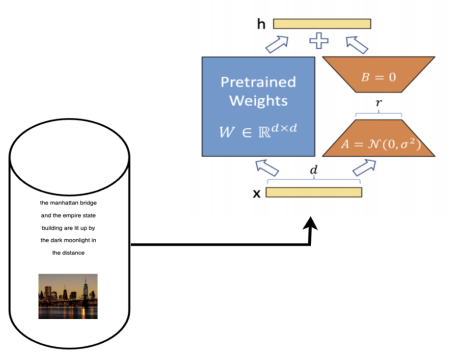

Figure 7. Model Training: CLIP / LoRA set up

The preliminary results from our training phase have been encouraging, showing improved alignment between the generated images and their corresponding captions, as evidenced by the CLIP scores. As we continue to refine our approach and expand the dataset, we aim to further enhance the model's performance and explore additional applications of our methodology in more complex VL tasks.

6. Results

We conducted an in-depth analysis of our model's performance, focusing on both quantitative benchmarks and quality evaluation of our dataset.

6.1. Quality Evaluation of the Dataset

In figures 4 and 5, we plotted the clip scores comparison distribution among generated images and original images. Both are done on a sample of 5,000 images. The plots have a similar trend indicating that the generated images align better with text and are more faithful representations. We see the the results are similar for both. Therefore, in line with our hypothesis the generated images shows better alignment with texts and are in turn more faithful.

6.2. Quantitative Evaluation

We trained our model on 100 epochs with a rank of 4 for the LoRA mechanism utilized in our model. In the table 1 we benchmarked our trained CLIP model with the Object localization and size which is a subset of (Zhao et al., 2023). Our model exhibits significant improvements over both the original CLIP model and the CLIP model fine-tuned with LoRA, achieving an accuracy of 92.89%. This enhancement underscores the effectiveness of our synthetic dataset in enhancing the model's understanding of objects within images.

We evaluated the performance of our model on the Object Localization and Size Task. Table 1 shows the results of our experiments, where we compared the performance of different models. Our baseline CLIP model achieved an accuracy of 81.58%, while the CLIP model fine-tuned with LoRA slightly decreased to 80.93%. However, when incorporating our synthetic data, the accuracy significantly improved to 92.89%. This remarkable enhancement demonstrates the effectiveness of our synthetic dataset in improving the model's understanding of objects within images.

Model	Arch	Object
CLIP	ViT-B/32	81.58%
CLIP + LoRA	ViT-B/32	80.93%
CLIP + LoRA + Our Synthetic Data	ViT-B/32	92.89%

Table 1. Performance Comparison on Object Localization and Size Task

Our results reveal some key insights into the performance gains achieved by our model. First, the incorporation of the LoRA mechanism results in a slight decrease in performance compared to the baseline CLIP model. This unexpected outcome suggests that further investigation into the interaction between the LoRA mechanism and the CLIP architecture may be warranted. However, the subsequent inclusion of our synthetic dataset significantly enhances performance, highlighting the importance of dataset quality in improving model effectiveness.

Overall, our results underscore the importance of both architectural enhancements and dataset quality in improving the performance of multimodal models. Further research into the design of attention mechanisms and the creation of diverse and representative datasets will be critical for advancing the state-of-the-art in visual understanding tasks.

7. Conclusion

In conclusion, the research presented in this report has successfully addressed the prevailing asymmetry in Vision-Language Model (VLM) development by focusing on enhancing the image modality. Through a novel methodology of generating faithful image-text pairs from existing captions, we have not only improved the alignment accuracy of these models but have also pioneered a more balanced approach to multimodal learning. This approach is anticipated to pave the way for more holistic and nuanced interpretations of visual and textual data, significantly advancing the field of VLMs.

Our findings indicate that by enriching the image representation space and creating high-quality image-text alignment data, VLMs can achieve a better understanding of the visual world, which is crucial for tasks like image captioning and visual question answering. The use of generated image-text

pairs has demonstrated a promising increase in model comprehension and performance, underscoring the potential of our innovative methods.

As we move forward, this research sets a precedent for future investigations into the integration of richer visual content within VLMs. By continuing to explore and refine these methodologies, the field can look forward to developing more sophisticated models that are capable of more accurately interpreting and interacting with the world around them. This balance of image and text modalities could revolutionize the way artificial intelligence understands and operates within our visual and linguistic world.

8. Contributions

In our team, each member played a crucial role in different aspects of our project. For most of the tasks that included research, thinking and discussions, the team worked together for it. When came the time to implement the ideas will split the tasks equally, according to the strength of each member. Aayush took charge of the training and benchmarking tasks, ensuring that our models were optimized and evaluated effectively. Milosh led the development of the synthetic dataset pipeline, streamlining the data generation process for our experiments. Paul conducted the empirical study, analyzing the results and drawing insights from our experiments. Carlone handled the preprocessing stage, ensuring that our data was cleaned and prepared appropriately for further analysis. Throughout the project, we held weekly meetings where everyone contributed their expertise and collaborated on various tasks, leveraging each other's strengths and skills to achieve our goals effectively.

References

Azizi, S., K. S. S. C. N. M. . F. D. J. Synthetic data from diffusion models improves imagenet classification, 2023.

Doveh, S., Arbelle, A., Harary, S., Herzig, R., Kim, D., Cascante-bonilla, P., Alfassy, A., Panda, R., Giryes, R., Feris, R., Ullman, S., and Karlinsky, L. Dense and aligned captions (dac) promote compositional reasoning in vl models, 2023a.

Doveh, S. et al. Text augmentation for image-text datasets. *arXiv preprint arXiv:2301.00567*, 2023b.

Dunlap, L., U. A. Z. H. Y. J. G. J. E. . D. T. Diversify your vision datasets with automatic diffusion-based augmentation, 2023.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. Jia–et al., 2021.

OpenAI. Gpt-4 technical report, 2023.

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. *arXiv preprint arXiv:2103.00020*, 2021.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2022.

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., and Lample, G. Llama: Open and efficient foundation language models, 2023.

Yang, L., X. X. K. B. S. Y. . Z. H. Freemask: Synthetic images with dense annotations make stronger segmentation models, 2023.

Yu, Z., Zhu, C., Culatana, S., Krishnamoorthi, R., Xiao, F., and Lee, Y. J. Diversify, don't fine-tune: Scaling up visual recognition training with synthetic images, 2023.

Zhao, T., Zhang, T., Zhu, M., Shen, H., Lee, K., Lu, X., and Yin, J. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations, 2023.

A. Github Repository

Here is the link for our Github repository with the code for our project: https://github.com/theAayushbajaj/self_align_clip.