Visual Language Alignment Tuning

Le Zhang, Qian Yang, Aishwarya Agrawal
Mila - Quebec AI Institute
Université de Montréal
le.zhang@mila.quebec

Abstract

Foundation models like CLIP are pivotal for advancing research in vision-language learning, as they simultaneously learn modality-specific representations and cross-modal alignment. However, training these models is resource-intensive, requiring hundreds of millions of image-text pairs and hundreds of GPUs, creating a barrier for advancing research on multimodal alignment. In this paper, we introduce the **S**wift **A**lignment of **I**mage and **L**anguage (SAIL) framework, which focuses on vision-language alignment by tuning a lightweight alignment layer added on top of frozen pretrained single-modality models. SAIL drastically reduces computational demands, requiring only a single GPU to align the pretrained feature spaces.

1 Introduction

Foundation vision-language models such as CLIP Radford et al. [2021] serve as a cornerstone in the vision-language research domain. These models are typically trained using contrastive learning on large-scale, noisy, web-collected image-text datasets, requiring massive computational resources, often involving hundreds of GPUs due to the large batch sizes necessary for contrastive learning. While these pretrained models have achieved significant success in computer vision research, the prohibitive computational demands hinder further exploration and innovation, particularly for academic researchers with limited access to these resources.

At their core, contrastive-based vision-language models aim to achieve two objectives simultaneously during the pretraining process: *modality-specific representation learning* and *multimodal representation alignment*. However, these two objectives—representation learning and representation alignment—are deeply intertwined, making it difficult to disentangle the model's alignment behavior and to independently improve each component. Moreover, the reliance on vast quantities of noisy, web-sourced image-text data poses a growing challenge, as curating high-quality datasets becomes increasingly difficult.

On the other hand, the computer vision (CV) and natural language processing (NLP) communities have developed models that already excel at learning modality-specific representations such as DINO family Caron et al. [2021], Oquab et al. [2023] for visual understanding and GTE family for text understanding Li et al. [2023c], Zhang et al. [2024]. Some studies further suggest that deep networks are converging towards a universal "statistical model of reality," capturing real-world structures across diverse domains and modalities Huh et al. [2024]. This convergence presents an opportunity to refine multimodal alignment by leveraging existing pretrained models without the need for retraining large-scale architectures from scratch. However, achieving alignment between pretrained vision and language models remains a non-trivial task. Prior studies, such as Zhai et al. [2022], have primarily focused on fine-tuning one modality's encoder while keeping the other frozen during the alignment process. While this approach reduces computational costs, it often necessitates training at least one encoder from scratch, which still demands significant resources.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

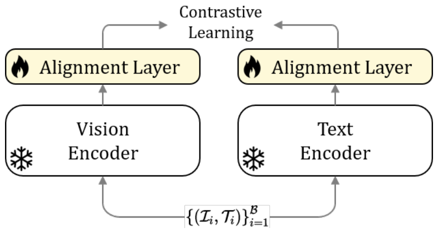

Figure 1: Overview of our method. Only light alignment layer is tuned during training process.

Motivated by these observations, we introduce **S**wift **A**lignment of **I**mage and **L**anguage (SAIL) framework, which decouples modality-specific representation learning from the alignment process. By utilizing pretrained unimodal vision and language models and introducing a lightweight mapping layer, we can achieve competitive multimodal alignment with a fraction of the computational resources (compared to learning the multimodal alignment from scratch) as shown in fig. 1. Specifically, SAIL enables the training of vision-language models on a single GPU in just one day, while maintaining performance comparable to models like CLIP. Our extensive experiments demonstrate the efficacy of this approach, showing that solely tuning a lightweight mapping network is sufficient to achieve robust multimodal representation alignment.

With SAIL's fast training, we are now able to thoroughly analyze and improve the alignment behavior of pretrained vision-language models. First, we investigate the design of the alignment network. Although prior studies on multimodal models suggest that simple MLPs work well as connectors Li et al. [2023b], we find that gated MLPs Shazeer [2020] consistently deliver superior results in our experiments. We also explore contrastive loss functions and discover that sigmoid loss, as used in SigLIP Zhai et al. [2023], consistently outperforms the InfoNCE loss used in CLIP Radford et al. [2021]. Additionally, treating negative pairs with equal weight to positive pairs improves the alignment process. Furthermore, SAIL enables us to leverage strong large language models that better understand complex sentences and support longer text inputs, overcoming the 77-token limit present in CLIP, thereby enhancing model performance in tasks involving richer textual data such as complex reasoning and long caption retrieval tasks.

Finally, we explore several underexamined aspects of vision-language learning. One key area we investigate is the trade-off between contrastive learning and batch size. While larger batch sizes are generally expected to enhance performance, we find that when the batch size exceeds 32k with sigmoid loss Zhai et al. [2023], training becomes unstable and prone to collapse. Our analysis attributes this instability to the temperature parameter, which controls the balance of gradients between positive and negative pairs. As the batch size increases, a higher temperature is needed to prevent negative pairs from overwhelming the learning process. Our results demonstrate that these factors are crucial to improving alignment, and we also observe that larger models are more easily aligned, consistent with the findings of Huh et al. [2024].

Bringing together all the findings described above, we demonstrate that alignment tuning effectively aligns the vision and language feature spaces and can be directly transferred to downstream zero-shot tasks, such as image classification, retrieval, and segmentation.

2 Related Work

2.1 Vision-Language Pretraining

Vision-language pretraining has significantly advanced multimodal understanding. Models like CLIP Radford et al. [2021], ALIGN Li et al. [2021], and SigLIP Zhai et al. [2023] leverage large-scale, noisy web datasets to train joint visual-text embeddings via contrastive learning, achieving robust performance across multiple tasks. However, their reliance on massive datasets and large batch sizes requires substantial computational resources, often involving hundreds of GPUs for weeks, which limits accessibility for many researchers.

BLIP Li et al. [2023a] aims to enhance pretraining efficiency and data quality but still demands considerable GPU resources and extensive data curation. The main challenge with these methods is the balance between the amount of computational investment and the quality of learnt multimodal representations, posing significant barriers for broader research communities.

2.2 Effective Vision-Language Alignment

Recent research has focused on efficiently aligning modality-specific feature spaces without retraining from scratch. Models like Flamingo Alayrac et al. [2022], LLaVA Li et al. [2023b], and LiT Zhai et al. [2022] leverage pretrained vision and language encoders, using simple strategies for alignment rather than costly retraining. These methods demonstrate that effective cross-modal alignment can be achieved with high-quality data and straightforward alignment mechanisms.

Flamingo and LLaVA use frozen vision encoders with minimal learnable parameters or simple MLPs to align modalities for image conditioned language modelling task, achieving strong performance with reduced resources. LiT freezes vision encoders and trains only the language component, allowing zero-shot transfer but failing to produce language-compatible visual features, limiting effectiveness in downstream tasks requiring deeply aligned multimodal representations.

In contrast, we aim to learn tightly aligned visual-text representations that support a wide range of downstream applications. By using lightweight mapping networks to effectively align pretrained models, SAIL achieves robust multimodal integration without extensive retraining, providing a scalable, resource-efficient solution accessible to researchers with limited computational resources.

3 Methods

SAIL consists of three main components: (1) the design of the alignment layer; (2) an effective loss function for learning a shared visual-language feature space; and (3) the use of small yet high-quality datasets.

Alignment Layer The alignment layer is crucial for bridging the feature spaces of frozen vision and language encoders. While LLaVA demonstrates success with simple MLPs, we empirically find that more sophisticated layers yield better alignment. Inspired by the effective use of Gated Linear Units (GLUs) Shazeer [2020] in large language models, we adopt a single-layer GLU for our alignment layer. As shown in table 1, the GLU offers significantly improved performance without increasing the computational burden (i.e., FLOPS) during forward passes, since the tunable parameters are minimal and restricted to the alignment layer.

Layer	# TP	FLOPs	INet-1k
Linear	3.15M	6.3 M	39.4
MLP	33.57M	67.1 M	51.2
GLU	54.6M	109.1 M	52.4

Table 1: Comparison of alignment layers using DINOv2 as the vision encoder and GTE-EN-LARGE as the text encoder trained on CC3M. TP is total trainable parameters, including 2 alignment layers. The input dimensionality is 2048 for visual features and 1024 for textual features, with an intermediate dimension of 4 times of input dimension and output dimension of 1024.

Contrastive Loss For contrastive learning, given a batch of image-text pairs $\{(\mathcal{I}_i, \mathcal{T}_i)\}_{i=1}^{\mathcal{B}}$ sampled from a dataset, and corresponding image and text encoders $\mathcal{F}_I$ and $\mathcal{F}_T$ as well as their corresponding alignment layer $\mathcal{G}_I$ and $\mathcal{G}_T$, we adopt the SigLIP approach, which uses a binary classification-based loss instead of the InfoNCE used in CLIP. By removing the need for heavy normalization in softmax operation, SigLIP has demonstrated improved efficiency and performance:

$$\mathcal{L} = -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\log\frac{1}{1+e^{z_{ij}(-t\mathbf{x}_i\cdot\mathbf{y}_j+b)}} = -\frac{1}{|\mathcal{B}|}\left(\underbrace{\sum_{i=1}^{|\mathcal{B}|}\log\sigma(-s_{i,i})}_{Alignment}+\underbrace{\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1,j\neq i}^{|\mathcal{B}|}\log\sigma(s_{i,j})}_{Uniformality}\right), \tag{1}$$

where $x_i = \mathcal{G}_I(\mathcal{F}_I(\mathcal{I}_i))$, $y_i = \mathcal{G}_T(\mathcal{F}_T(\mathcal{T}_i))$. Their L2-normalized features are denoted as $\mathbf{x}_i = \frac{x_i}{\|x_i\|_2}$ and similarly $\mathbf{y}_j = \frac{y_j}{\|y_j\|_2}$. The similarity score is defined as $s_{ij} = -t\mathbf{x}_i \cdot \mathbf{y}_j + b$. Here, t and b represent temperature scaling and bias, respectively, while z_{ij} is 1 if $i = j$, and -1 otherwise. We find it beneficial to treat all pairs equally by normalizing over $|\mathcal{B}|^2$ instead of $|\mathcal{B}|$, ensuring each pair contributes uniformly, which helps in robust alignment.

High-Quality Data and Multiple Positive Caption Contrast Recent work suggests that effective vision-language models can be trained on smaller but higher-quality datasets Li et al. [2023a,b], Zheng et al. [2024], Fan et al. [2024]. To collect high-quality image-caption pairs, we leverage multimodal large language models (MLLMs) for recaptioning existing datasets. Methods like DreamLip use MLLMs such as ShareGPT4 Chen et al. [2023a], LLaVA Li et al. [2023b], and InstructBLIP Dai et al. [2023] to enhance datasets like CC3M, CC12M, and YFCC15M, generating richer captions. We observe that longer captions benefit retrieval tasks requiring deeper semantic understanding, whereas shorter captions are advantageous for image classification due to their concise focus. To leverage both advantages, we propose mixing long and short captions within each training batch, providing diverse training signals for learning robust representations.

Given an image-text batch $\{(\mathcal{I}_i, \mathcal{T}_i)\}_{i=1}^{\mathcal{B}}$ and corresponding high-quality captions $\mathcal{T}_i^{HQ}$ generated by ShareGPT4, the multiple positive caption contrast loss is defined as:

$$\mathcal{L} = -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\left(\log\frac{1}{1+e^{z_{ij}(-t\mathbf{x}_i\cdot\mathbf{y}_j+b)}} + \log\frac{1}{1+e^{z_{ij}(-t\mathbf{x}_i\cdot\tilde{\mathbf{y}}_j+b)}}\right), \tag{2}$$

where $\tilde{\mathbf{y}}_j$ is the L2 normalized representation of the high-quality caption. This approach effectively combines information from both standard and enriched textual descriptions, improving the model's robustness in downstream tasks that require nuanced cross-modal understanding.

4 Experiments

We use DINOv2-ViT-L Oquab et al. [2023] as the vision encoder and GTE-en-large-v2 Li et al. [2023c] as the text encoder. Since SAIL only tunes the alignment layer, we efficiently perform contrastive learning with a batch size of 36k using a single A100 for 3 hours. Our training setup includes a learning rate of 1e-5, weight decay of 1e-7, a cosine learning rate scheduler, and the LION optimizer Chen et al. [2023b] with beta1 = 0.9 and beta2 = 0.99. We empirically set the temperature (t) to 20 and bias (b) to -10. For evaluation, we perform zero-shot image classification on ImageNet-1k, zero-shot retrieval on COCO, and leverage MaskCLIP Zhou et al. [2022] for zero-shot segmentation on ADE20K Zhou et al. [2017] and VOC20 Everingham et al. [2015]. We also assess complex visual reasoning on Winoground Thrush et al. [2022].

We compare SAIL with CLIP-L, trained on 400M image-text pairs, as summarized in table 2. Despite using only 7M pairs, SAIL performs well on classification and retrieval tasks, demonstrating the effectiveness of lightweight alignment layers between frozen feature spaces. This approach preserves the strengths of pretrained models, enabling SAIL to significantly outperform CLIP in zero-shot segmentation and leverage pretrained features effectively. Additionally, SAIL surpasses CLIP on complex reasoning tasks (Winoground), benefiting from its stronger text encoder trained on more diverse and complex data.

Model	Trainable Params	Image-Text Data	Training GPU Hours	ZS-Classification	ZS-Retrieval		ZS-Segmentation		Winoground		
				ImageNet-1k	COCO I2T	COCO T2I	ADE20K	VOC 21	Text	Image	Group
CLIP-L	428M	400M	6144	76.2	56.3	37.7	1.2	41.8	28.5	11.25	8.25
SAIL	54.6M	2.2M	2	52.4	44.9	33.5	11.6	68.4	31	10.75	8.75
SAIL	54.6M	7.4M	2	61.2	45.2	33.4	13.8	66.4	33.25	13	11
SAIL	54.6M	12.9M	3	61.6	52.2	40.8			29	12.5	7.75

Table 2: Zero-shot Evaluation: ImageNet results are reported as accuracy, retrieval performance as Recall@1, and segmentation as mIOU. Winoground is reported following original metrics.

5 Conclusion

This paper introduces visual-language alignment tuning, where a lightweight MLP layer is fine-tuned over frozen pretrained unimodal models using small yet high-quality datasets. This approach achieves strong transfer performance on zero-shot vision-language tasks. Furthermore, we find that alignment tuning not only facilitates effective cross-modal integration but also retains the strengths of the pretrained models, such as DINO's robust object-level features and the ability of language models to handle long and complex sentences.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023a.

X Chen, C Liang, D Huang, E Real, K Wang, Y Liu, H Pham, X Dong, T Luong, CJ Hsieh, et al. Symbolic discovery of optimization algorithms. arxiv 2023. *arXiv preprint arXiv:2302.06675*, 2023b.

Wenliang Dai, Junnan Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*, 2, 2023.

Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36, 2024.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023a.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. MΘ3 it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023b.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023c.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18123–18133, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *arXiv preprint arXiv:2407.19669*, 2024.

Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. Dreamlip: Language-image pre-training with long captions. *arXiv preprint arXiv:2403.17007*, 2024.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022.