Assessing and Learning Alignment of Unimodal Vision and Language Models

Supplementary Material

A. Reproducibility Statement

To ensure the reproducibility of our work, we are committed to making all training code, datasets, and model weights publicly available at Project Page. Detailed documentation will accompany the codebase to facilitate easy replication of our experiments. Hyperparameter settings, training configurations, and any preprocessing steps will also be thoroughly outlined. By providing these resources, we aim to promote transparency, enable future research, and support the broader community in building upon our work.

B. Alignment Assessment Training Details

Alignment Probing The alignment probing method uses contrastive learning to train linear layers, referred to as alignment layers, for aligning pretrained unimodal vision and language representation spaces.

Specifically, with a **frozen** image encoder $\mathcal{F}_I(\cdot)$ and a **frozen** text encoder $\mathcal{F}_T(\cdot)$, the corresponding **linear** layers $\mathcal{G}_I(\cdot)$ and $\mathcal{G}_T(\cdot)$ are trained using the refined sigmoid loss on CC3M dataset with ShareGPT4-enhanced captions and incorporating the multiple positive caption contrast, as described in Sec. 3.1. For optimization, we use the LION optimizer (with $\beta_1 = 0.9$, $\beta_2 = 0.99$), a learning rate of 10^{-5}, and a weight decay of 10^{-7}. We use a temperature of $t = \log 20$ and a bias of $b = -10$. The output dimensionality of the linear layer (alignment dimensionality) is 2048. Training runs for 100 epochs with a batch size of 32,768, using a fixed image resolution of 224.

Linear Probing vs. Alignment scores Fig. 8 illustrates the relationship between our alignment metric and the ImageNet linear probing classification accuracy of models. Compared to kNN (refer to Sec. 1), the linear correlation between these two metrics is weaker, with a Pearson correlation coefficient of 0.847. This highlights that nonlinear separability (i.e., clustering quality) matters more than linear-separability for image-text alignment.

C. Additional Comparison with ShareLock

In Sec. 3.2.1 and Sec. 3.2.2, we compare our method directly with the concurrent work ShareLock, using the reported results from [37], as the code was not open-source at the time of submission. ShareLock utilizes the LLaMA3-8B as the language encoder, which differs from the NV-Embed-2 language encoder used in SAIL. To ensure a fair comparison, we reproduced ShareLock's results after consulting the authors, using the same vision and language backbones as SAIL (DINOv2-B and NV-Embed-2). We adhered strictly to the training details provided in the original paper [37] and present evaluation results for classification and retrieval tasks in Tab. 6.

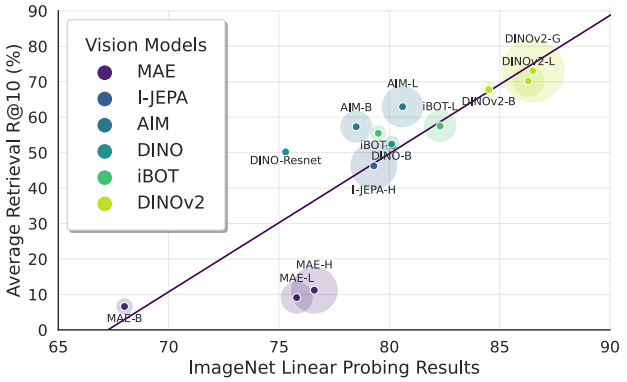

Figure 8. **Linear alignment probing results** between Imagenet linear probing accuracy and average retrieval R@10 (our metric). MAE serves as an outlier, achieving high linear probe performance but low alignment performance.

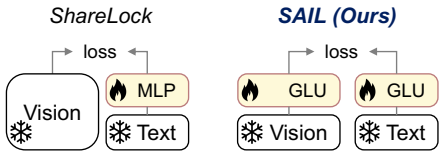

Figure 9. **Method comparison.** SAIL shows consistent improved performance over ShareLock.

The ShareLock results demonstrates that replacing LLaMA3-8B with NV-Embed-2 significantly improves alignment performance across benchmarks. Also we see that, using the same vision and language backbones, SAIL (NV2) consistently outperforms ShareLock (NV2) across all tasks by a significant margin. This highlights the effectiveness of incorporating alignment layers for both vision and language models (see Fig. 9 for differences), as well as the advantages of our proposed optimized training methodologies.

Experiments without enhanced caption We provide additional results in Tab. 7 by training SAIL on CC12M raw dataset without long captions. We see that SAIL still significantly outperforms both ShareLock and LiT across all benchmarks, further validating the effectiveness of our method.

Data	Model	MSCOCO I2T	MSCOCO T2I	Flickr30k I2T	Flickr30k T2I	Winoground T.	Winoground I.	Winoground G.	MMVP 10 Avg.	ImageNet Top1.	10 Classification Avg.
Model Architecture: ViT-B/16											
CC12M	DreamLIP	53.3	41.2	82.3	66.6	26.0	10.00	7.25	24.0	50.3	49.9
	LiT‡	30.0	16.5	54.8	38.5	24.3	6.5	4.8	-	56.2	-
	ShareLock(Llama3)†‡	26.0	13.5	53.9	34.9	26.3	12.8	5.3	-	59.1	-
	ShareLock(NV2)†	39.6	23.1	68.1	49.3	33.25	13	9.75	15.56	61.9	62.0
	SAIL-B (GTE)†	48.2	37.9	76.5	63.9	31.0	11.5	9.5	23.0	58.7	57.7
	SAIL-B (NV2)†	57.3	45.3	84.1	70.1	35.0	17.25	13.0	24.4	68.1	65.4
LAION400M	CLIP-B	55.4	38.3	83.2	65.5	25.7	11.5	7.75	19.3	67	65.5
Model Architecture: ViT-L											
23M Merged	SAIL-L (NV2)†	62.4	48.6	87.6	75.7	40.25	18.75	15.0	28.9	73.4	72.1
LAION400M	CLIP-L	59.7	43.0	87.6	70.2	30.5	11.5	8.75	20.0	72.7	75.9

Table 6. **Results** on standard retrieval, complex reasoning, visual-centric, and classification tasks. We report Recall@1 for MSCOCO and Flickr30k, Text, Image, and Group scores for Winoground, and the average score across 9 visual patterns for MMVP. ‡ indicates cited results, and † denotes a ViT patch size of 14. 10 Classification tasks include: Food101, CIFAR10, CIFAR100, SUN397, Cars, Aircraft, DTD, Pets, Caltech101, and Flowers.

Data	Model (DINOv2-B)	MSCOCO I2T	MSCOCO T2I	Flickr30k I2T	Flickr30k T2I	Winoground T.	Winoground I.	Winoground G.	MMVP 10 Avg.	ImageNet Top1.	10 Classification Avg.
CC12M raw	LiT‡	30.0	16.5	54.8	38.5	24.3	6.5	4.8	-	56.2	-
CC12M raw	ShareLock (NV2)	39.6	23.1	68.1	49.3	33.25	13	9.75	15.56	61.9	62.0
CC12M raw	SAIL-B (NV2)	45.6	32.9	74.2	60.6	35.0	19.0	14.25	25.2	69.2	66.4

Table 7. Trained on CC12M raw captions.

D. Dataset used for evaluation

(a) some plants surrounding a lightbulb

(b) a lightbulb surrounding some plants

Figure 10. An example from Winoground.

Winoground evaluation Winoground [40] is a benchmark designed to evaluate the ability of vision-language models to perform visio-linguistic compositional reasoning, we provide one example as in Fig. 10. The task involves matching the correct image-captions pairs given two images and two captions, where the captions contain identical sets of words but in different orders, requiring fine-grained reasoning about the visual and textual alignment.

Performance is measured using three metrics: text score, image score, and group score, defined as follows. Given two image-text pairs (I_0, T_0) and (I_1, T_1), and a similarity function $s(\cdot)$ provided by the model:

The **text score** evaluates if the ground-truth caption for each image is scored higher than the alternative caption. It is computed as:

$$f(T_0, I_0, T_1, I_1) = \begin{cases} 1 & \text{if } s(T_0, I_0) > s(T_1, I_0) \\ & \text{and } s(T_1, I_1) > s(T_0, I_1) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

The **image score** tests whether the correct image is selected for each caption. It is computed as:

$$g(T_0, I_0, T_1, I_1) = \begin{cases} 1 & \text{if } s(T_0, I_0) > s(T_0, I_1) \\ & \text{and } s(T_1, I_1) > s(T_1, I_0) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

The **group score** combines the two previous metrics, requiring both to be correct simultaneously:

$$h(T_0, I_0, T_1, I_1) = \begin{cases} 1 & \text{if } f(T_0, I_0, T_1, I_1) \\ & \text{and } g(T_0, I_0, T_1, I_1) \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

These metrics collectively assess whether the model can align text and images accurately while reasoning over compositional semantics.

MLLM benchmarks In Sec. 3.2.4, we combine SAIL vision encoder with LLaVA-1.5 and evaluated on various downstream VQA and instruction-following benchmarks. Below we provide a description of each of these benchmarks.

- **SEED** [23]: SEED-Bench offers a comprehensive evaluation framework with **19K multiple-choice questions**, featuring accurate human annotations—six times larger than existing benchmarks. It spans **12 evaluation dimensions**, covering comprehension in both **image** and **video modalities**. The use of multiple-choice questions with human-annotated ground truth answers ensures **objective and efficient model assessment**, removing the need for human or GPT intervention during evaluation.
- **GQA** [16]: GQA stands out as a **dataset for real-world visual reasoning** and compositional question answering, addressing key limitations of earlier VQA datasets. It emphasizes **reasoning, compositionality**, and the **grammar-based generation** of natural language queries, pushing models to engage in structured and logical visual understanding.
- **VizWiz** [12]: VizWiz features over **31,000 visual questions** originating from visually impaired individuals who used mobile phones to capture images and record spoken queries. Each question is paired with **10 crowdsourced answers**, introducing challenges such as **blurry images, partial scenes**, and diverse visual content, providing a real-world perspective on VQA.
- **PoPE** [25]: PoPE targets **Object Hallucination** in multimodal large language models (MLLMs) by focusing on challenging visual reasoning tasks. It transforms hallucination evaluation into a **binary classification task**, using **Yes-or-No questions** about specific objects (e.g., "Is there a car in the image?"), offering a direct and interpretable measure of model accuracy in visual interpretation.
- **TextVQA** [39]: TextVQA challenges models to **extract and reason about textual information** embedded in images, such as names, prices, and other details. It heavily relies on **Optical Character Recognition (OCR)** to parse diverse and complex text inputs. The dataset pushes OCR systems to handle variations in **font styles, sizes, orientations**, and noisy scenes, providing critical inputs for downstream reasoning tasks.
- **MMBench** [29]: MMBench is a **systematically designed benchmark** for evaluating the diverse abilities of large vision-language models (VLMs). It includes **3,000+ multiple-choice questions** across **20 ability dimensions**, such as **object localization** and **social reasoning**. Each dimension is represented by **125+ balanced questions**, ensuring robust evaluation. Tasks such as text interpretation within images further emphasize the importance of **OCR capabilities** in vision-language modeling.
- **VQAv2** [11]: As one of the most widely used benchmarks for VQA, VQAv2 introduces **balanced questions** to mitigate language biases. It emphasizes **visual reasoning**, requiring models to align language understanding with accurate visual grounding, setting a strong standard for comprehensive VQA tasks.

E. Pre-encoding Efficiency

The SAIL training pipeline comprises two key stages: pre-encoding and alignment tuning. Here, we provide an estimate of the pre-encoding speed for models used in constructing SAIL. Encoding speed is influenced by factors such as hardware capabilities, model architecture, and the availability of acceleration techniques like FlashAttention. Additionally, for language models, sentence length significantly affects encoding performance.

Since encoding times depend on hardware and model configurations, we report approximate times based on our training setup, utilizing a single A100-80G GPU:

- With DINOv2-L using scaled dot-product attention, encoding 224x224 resolution images from CC3M achieves a throughput of approximately $\sim$ 830 samples/s.
- For GTE-en-large-v1.5 with FlashAttention, the throughput is $\sim$ 2350 samples/s for short raw captions and $\sim$ 130 samples/s for longer, high-quality captions (truncated to a maximum of 1024 tokens).
- With NV-embed-2, the throughput is $\sim$ 170 samples/s for short raw captions and $\sim$ 25 samples/s for longer, high-quality captions (truncated to a maximum of 1024 tokens).

With acceleration methods such as FlashAttention and vLLM, the encoding speed could be further enhanced for these models. Note that encoding is performed only once and reused multiple times during training.