Deciphering Cross-Modal Alignment in Large Vision-Language Models via Modality Integration Rate

Supplementary Material

1. Related Work

1.1. Vision-Language Foundation Model.

Vision-Language Models (VLMs) have emerged as a significant advancement in nowadays' multi-modal learning, capable of understanding and generating human-like responses based on visual and textual inputs. Early models like CLIP (Contrastive Language–Image Pre-training) [25] marks a pivotal moment by aligning images and text in a shared embedding space, enabling the strong cross-modal understanding. Following CLIP, models like BLIP (Bootstrapping Language-Image Pre-training) [15, 16] extends this foundation, enhancing the fusion of vision and language modalities by leveraging more complex pre-training objectives. As the capabilities of Large Language Models (LLMs) [28, 33] progressed, their integration with vision models gave rise to more powerful instruction-following Large Vision-Language Models (LVLMs) [2, 7, 9, 18–20, 24, 32, 34]. Early models such as Flamingo [1] and PaLM-E [10], and more recent ones like LLaVA [20] and Qwen-VL [2], exemplify this trend.

Most LVLMs share three essential components: *the vision encoder*, *the vision-language connector*, and *the language decoder*. *The vision encoder* is responsible for extracting precise features from images, capturing both detailed and abstract visual information. Popular choices include CLIP [25], OpenCLIP [12], EVA-CLIP [27], SigLIP [31] and DINO series [21], which are designed to provide both coarse-grained and fine-grained visual guidance. *The vision-language connector* plays a critical role in mapping the encoded visual features into a format that can be interpreted by the language model. Common designs include simple MLP projectors and the Q-Former used in BLIP-2, while more advanced solutions, such as the vision abstractor in mPLUG-Owl [30] and QLLaMA in Intern-VL [8], push the boundaries of cross-modal alignment. *The language decoder* is typically a pre-trained LLM designed to handle large-scale language data, ensuring that the model has robust instruction-following and conversational abilities. However, the central challenge in building a strong LVLM lies in bridging the modality gap between vision and language. The goal is to ensure that the language decoder can process visual tokens as naturally as it does language tokens, enabling smooth and meaningful conversations with multi-modal inputs. This crucial process is typically addressed during the pre-training stage of LVLM development. In this paper, we focus on evaluating and improving cross-modal alignment during the pre-training of LVLMs, a critical step in enhancing their overall performance and ensuring seamless interaction between visual and textual modalities.

1.2. Cross-Modal Alignment in LVLMs.

Cross-modal alignment plays a pivotal role in building a strong LVLM that can well support users to input images/videos and the model can understand the multi-modal contents. For the connector module of cross-modal alignment, there are typically three types widely used in current LVLMs: 1) Flamingo-style [1]. The perceiver resampler projects the vision features into the fixed number of vision tokens, and the language decoder captures the vision information by introducing cross-attention in Gated XATTN-DENSE layer. 2) BLIP-2-style [16]. A Q-Former to extract the instruction-aware information from vision tokens through cross-attention and pass the extracted tokens to the language decoder. 3) LLaVA-style [20]. A simple MLP projector directly map the vision tokens into the text embedding space.

Current Large Vision-Language Models (LVLMs) typically undergo a pre-training stage specifically designed for cross-modal alignment. As a result, the quality of the pre-training data and the strategies employed are critical for enhancing this alignment. Early datasets, such as COCO [17], Flickr30k [23], and LAION-400M [26], focus on short captions describing visual content. More recent datasets like ShareGPT4V [6] and ALLaVA [5] feature longer captions, aiming to provide richer descriptions to encourage the model to fully utilize the dense information of vision tokens. Besides, some works have shown that incorporating grounding information [22] or dense priors [14] in the captions further enhances LVLMs' ability to comprehend visual inputs. High-quality data plays a key role in improving the cross-modal alignment in LVLMs, driving advancements in multi-modal understanding. Various metrics or evaluation tools such as loss, perplexity, in-context evaluations and rank-based metrics [29] are transferred from LLMs to explore their potentials on quantifying LVLM pre-training, while somehow exhibiting the limitations under some particular LVLM pre-training settings.

#Data Scale	100K	200K	400K	600K	800K	1M	1.2M	1.4M	1.6M	1.8M
Average score on 7 popular multi-modal benchmarks										
Post-SFT Performance	56.5	58.1	59.5	59.9	60.0	59.6	60.1	59.8	60.0	59.8
MMD	0.535	0.535	0.535	0.535	0.535	0.535	0.535	0.535	0.535	0.535
MMD + Layer-Wise Accum.	0.201	0.204	0.203	0.208	0.209	0.210	0.214	0.207	0.214	0.210
KID	0.536	0.536	0.536	0.536	0.536	0.537	0.536	0.537	0.536	0.537
KID + Layer-Wise Accum.	0.264	0.268	0.266	0.271	0.272	0.274	0.277	0.270	0.277	0.274
Mutual Information	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Mutual Information + Layer-Wise Accum.	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
MIR w/o (a) (b)	6.645	6.056	4.984	3.443	4.591	NaN	3.883	NaN	NaN	NaN
MIR w/o (b)	6.645	5.964	4.921	NaN	NaN	NaN	NaN	NaN	NaN	NaN
MIR w/o (c)	5.041	4.477	3.572	3.235	3.180	3.119	3.173	3.115	3.161	3.191
MIR	5.041	4.477	3.576	3.245	3.192	3.133	3.187	3.131	3.176	3.206

Table 1. Comparison with conventional metrics and ablation studies on the the proposed enhancements. We validate the effectiveness of different metrics and settings based on the pre-training data scaling experiment of LLaVA-v1.5 7B model. The conventional metrics selected here includes MMD (Maximum Mean Discrepancy), KID (Kernel Inception Distance), and the mutual information. The components we ablate in MIR includes (a) Outlier Token Removal, (b) Text-Centric Normalization, and (c) Newton-Schulz Square Root Approximation. For vanilla MMD, KID, and Mutual Information, we report the result calculated at the input space of base LLM, as well as the layer-wise accumulated results.

2. Appendix Experiments

2.1. Comparison with Conventional Metrics

Here are detailed clarifications about why we refer Fréchet Distance as the basic metric component and why we propose Newton-Schulz based Square Root Approximation in MIR. The clarifications are expended upon the last paragraph in Sec. 2.1 of our paper, with further quantitative results on the effectiveness and efficiency of MIR.

Generally, MIR enjoys the following properties: 1) representative of the difference, 2) computationally efficient, and 3) flexible in terms of sample number. Given the typical discrepancy in the number of vision and text tokens, i.e., $r \neq s$, conventional metrics that require matching sample sizes between domains (*e.g.*, KL-divergence and CKA (Centered Kernel Alignment) [13]) are unsuitable for this scenario. Additionally, due to the high diversity and dimensionality of token features in LVLMs, MMD (Maximum Mean Discrepancy) [4] and KID (Kernel Inception Distance) [3] struggle to capture complex, high-level visual distributions, and the mutual information is often hindered by its high computational complexity. In contrast, Fréchet Distance [11] offers a more adaptive way with lower complexity for efficient domain divergence computation in LVLMs. However, it suffers from 1) the absolute value difference and abnormal value problem across different layers or models, leading to unfair comparison among them; 2) the computational cost of the matrix square root term in PyTorch, leading to the difficulty when computing on high-dimension matrices. Thereby, we enhance Fréchet Distance with our targeted improvements for multi-modal scenario, including Newton-Schulz based Square Root Approximation for efficient MIR computation.

We compare MIR to other conventional metrics with a comprehensive quantitative study. From Table 1, it is clear that the values of the conventional metrics such as MMD, KID, and mutual information are either irregular or reaching NaN, only MIR can well reflect the trend of model abilities as the pre-training data is scaling up. Table 1 also shows the Newton-Schulz Square Root Approximation brings no more than 1% error compare with traditional square root solving method in MIR computation.

2.2. Newton-Schulz Square Root Approximation

As we illustrated in Sec. 2.1 of our paper, for the efficiency, we compute the matrix square root term $(\Sigma_{v,k}\Sigma_{t,k})^{1/2}$ through Newton-Schulz iteration in the PyTorch version of MIR computation. Apparently, based on the Schur theorem, $\Sigma_{v,k}\Sigma_{t,k}$ is also positive semi-definite as the result of the positive semi-definite properties of covariance matrices $\Sigma_{v,k}$ and $\Sigma_{t,k}$. Therefore, we can define $\mathbf{A} = \Sigma_{v,k}\Sigma_{t,k}$ and $\mathbf{A}$ has its square root that can be denoted as $\mathbf{A}^{1/2}$.

Here we show the basic steps about how we compute the square root $\mathbf{A}^{1/2}$ via Newton-Schulz iteration. First, we need to initialize iterative terms like $\mathbf{Y}_0 = \mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{Z}_0 = \mathbf{I}$, where $\mathbf{I}$ is the $n \times n$ identity matrix. Then, we adopt the Newton-Schulz method to iteratively compute $\mathbf{Y}_{k+1}$ and $\mathbf{Z}_{k+1}$ respectively, *i.e.*,

$$\begin{aligned} \mathbf{Y}_{k+1} &= \frac{1}{2}\left(\mathbf{Y}_k + \mathbf{Z}_k^{-1}\mathbf{A}\right), \\ \mathbf{Z}_{k+1} &= \frac{1}{2}\left(\mathbf{Z}_k + \mathbf{Y}_{k+1}^{-1}\mathbf{A}\right). \end{aligned} \tag{1}$$

After T iterations, $\mathbf{Y}_T$ will be gradually close to the actual square root $\mathbf{A}^{1/2}$, *i.e.*, we can approximate it as

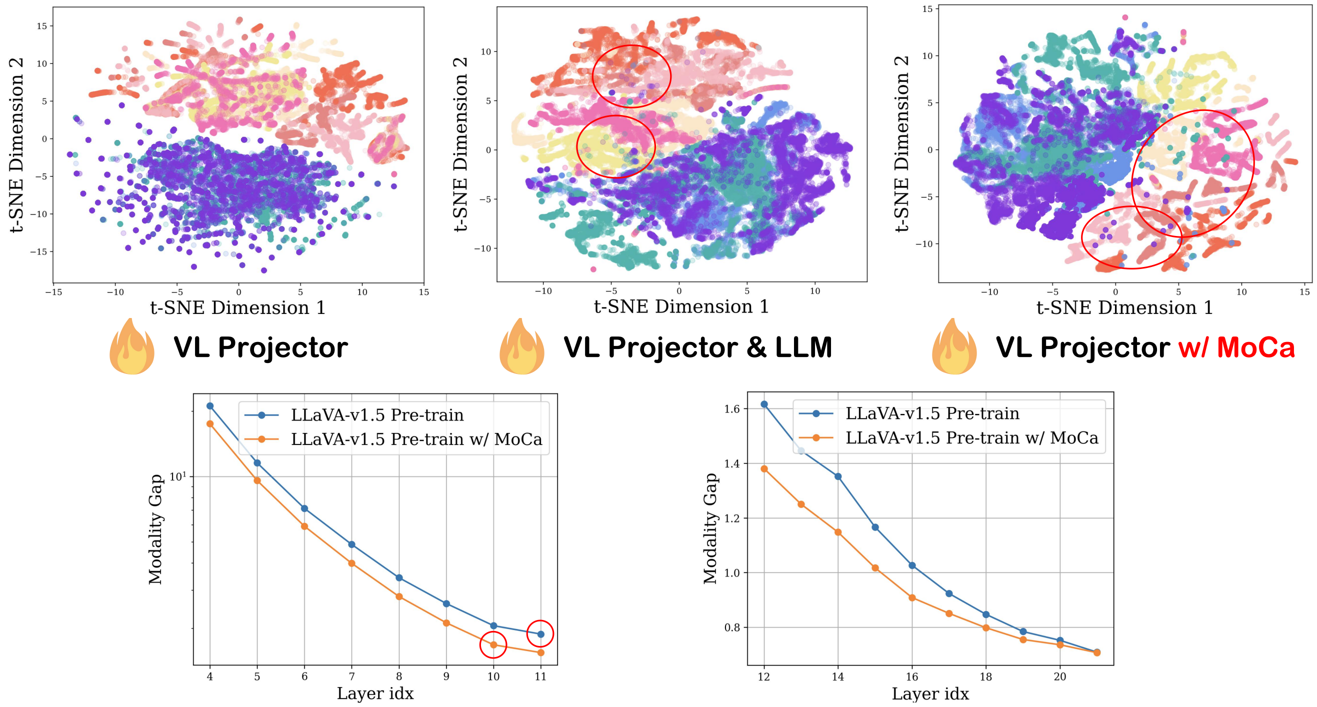

Figure 1. (**Top**) t-SNE ablation. (**Bottom**) layer-wise ablation.

$\mathbf{Y}_T \approx \mathbf{A}^{1/2}$. This approximation also supports the backward pass in the gradient computation. Suppose the square root obtained in the forward pass as $\mathbf{Y}$, the gradient of $\mathbf{A}$ can be updated as

$$\nabla \mathbf{A} = \nabla \mathbf{Y} \cdot \left(\mathbf{Y}^{-1}\right)^\top \cdot \left(\mathbf{Y}^{-1}\right)^\top . \qquad (2)$$

2.3. Analysis & Visualization for MoCa.

1) t-SNE results in Figure 1 shows that training only VL projector leads to a relatively larger modality gap, while incorporating MoCa or unlocking LLM significantly narrows this gap, promoting better modality space sharing (red circle). 2) Layer-wise ablation results in Figure 1 shows MoCa helps the model reduce the modality gap in each early layers. 3) The results below proves that MoCa can also improves pre-training quality in 13B model.

13B Model	MIR↓	Avg Acc	MMStar	MME	MMB	SEED^I	TQA
LLaVA-v1.5	2.546	61.1	32.3	1523.8	67.9	68.1	60.9
+MoCa	2.439	62.4	35.7	1530.2	68.9	68.7	62.1

Table 2. MoCa's effectiveness on 13B model

2.4. Time Efficiency of MIR

We select different numbers of data samples for exploring MIR's calculation efficiency, on a single NVIDIA A100-80G GPU. As shown in Table 3, MIR is generally efficient for evaluating a pre-trained LVLM. In most of cases, the MIR value is enough reliable when using more than 20 samples for computation (See Section 2.6).

#Samples	1	10	20	50	100
Time (s)	13.3	19.2	23.3	33.1	65.9

Table 3. MIR computation time cost when using different amount of image-text pair samples.

2.5. The Necessity of Text-Centric Normalization

The computation of our MIR requires text-centric normalization for both vision tokens and text tokens. This design ensures fairness in cross-layer comparisons of MIR, as FID values are sensitive to the absolute magnitudes of the inputs. Besides the ablation listed in Table 1, to explore this further, we ablate the scaling factor used in MIR computation, and the results are shown below:

Without text-centric normalization, the MIRs across different layers of the language model exhibit a pattern of first decreasing and then increasing, with the final MIR even higher than that of the first layer. This is counterintuitive because the deepest layer is closest to the language supervision, and the vision/text tokens at that layer should be more tightly aligned. For example, if we attempt to find the closest text embeddings for the vision tokens in the deepest layer across the vocabulary, we will observe much more se-

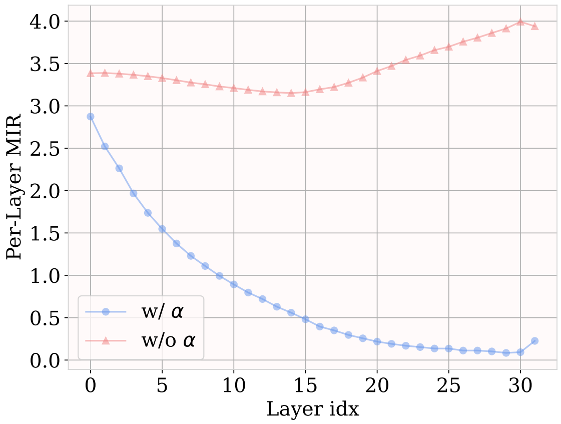

Figure 2. Text-centric normalization is necessary for MIR computation. We ablate the α in MIR and find that it can help MIR to realize the fair cross-layer comparison.

mantic alignment compared to the vision tokens in the first layer. Therefore, without text-centric normalization, MIRs across layers become incomparable due to differences in absolute values, rendering cross-layer MIR comparisons unfair. Hence, applying text-centric normalization in MIR is essential for meaningful comparisons.

2.6. Is MIR sensitive to the number of data sample?

As we clarified in the Method, we use 100 random selected images from TextVQA validation set and text data from CNN/DM for MIR calculation. Hereby, we explore the sensitivity of MIR to the number of data samples. We randomly choose 10 sets of the certain number of data samples to compute MIR for pre-trained LLaVA-v1.5 7B model, reporting the average values and ranges under different data sample numbers.

The results are as below:

#Samples	1	5	10	20	50
LLaVA-v1.5 7B	3.380	3.358	3.377	3.379	3.374
#Samples	100	200	500	800	1000
LLaVA-v1.5 7B	3.375	3.376	3.376	3.376	3.376

Table 4. The mean value of MIR gradually becomes stable with the increase of sample number.

It can be concluded that, if we use more than 20 samples to compute MIR, the fluctuation range is relatively small and we just need to compute MIR for one times as the negligible error, instead of computing for multiple times to get average value. Overall, MIR is relatively robust to the number of data samples, which is effective and reliable when $N \geq 20$.

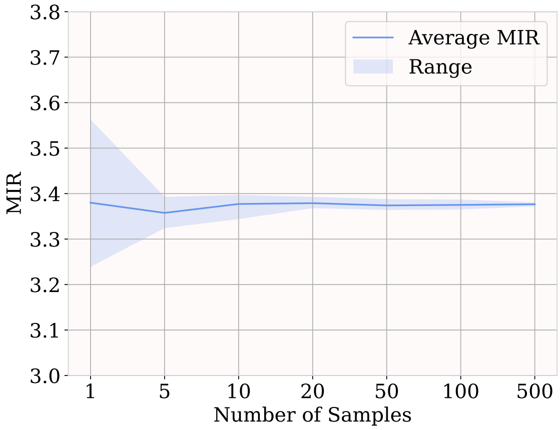

Figure 3. The fluctuation amplitude of MIR gradually decreases with the increase of sample number.

2.7. Further Discussion *w.r.t* PPL in LVLMs

In Figure 1 of our paper, we show the PPL is not precise to indicate the pre-training quality. This result is draw from computing PPL on LLaVA-v1.5 7B model that is pre-trained on GPT-style pre-training data (i.e., ALLaVA and ShareGPT4V-PT) and evaluating with the samples selected from ShareGPT4V, which means the training data and the evaluation samples are from the same domain. Here we should argue that PPL is much less reliable when the pre-training data has domain gap with the evaluation samples. To this end, we conduct the experiments on the ∼1.2M data by mixing LLaVA's BLIP-2-generated 558K data and ALLaVA, to pre-train LLaVA-v1.5 7B model with different scale of data. Then we follow the same evaluation settings to compute PPL and MIR, showcasing the results at Figure 4.

It indicates that PPL is not appropriate for evaluating the pre-training quality of LVLMs, which is struggling to deal with LVLMs' diverse pre-training data from multiple domains nowadays. In contrast, MIR offers a reliable evaluation for LVLM pre-training without SFT.

2.8. Larger LVLMs

We further study the MIRs of LVLMs that have different scale of base LLMs. All of pre-training data and recipes are the same with the official setting of LLaVA-v1.5. The results are listed in Table 5.

The results above show that the 13B base LLM achieves a lower MIR than the 7B base LLM, indicating that the larger, well-trained LLM has a stronger capability to narrow the modality gap in the shallow layers (as MIR is heavily influenced by the larger modality gap in the shallow layers of the language model). This is also consistent with the improved post-SFT multi-modal performance of the 13B

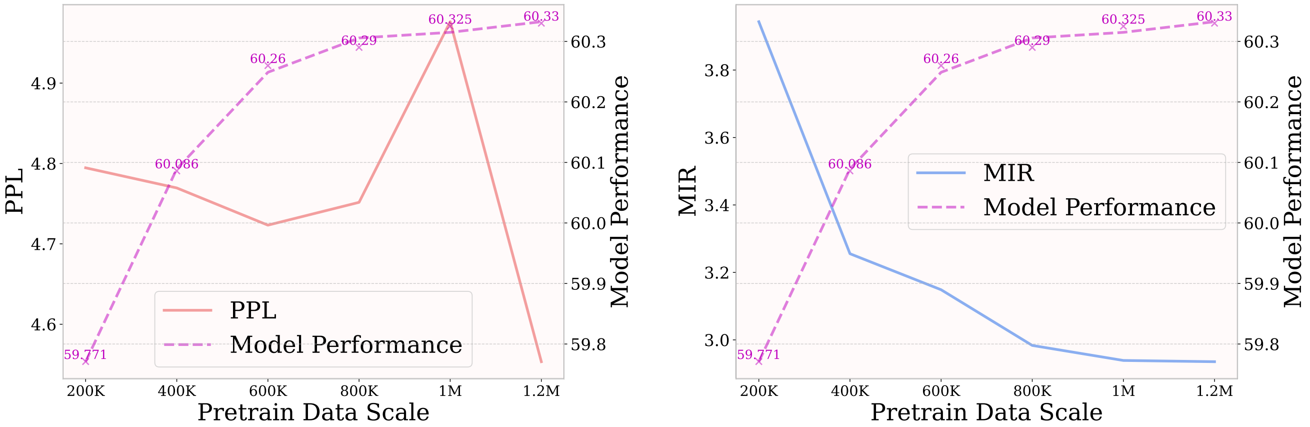

Figure 4. PPL is much less reliable when the pre-training data has domain gap with the evaluation samples.

Base LLM	Vision Encoder	Projector	Pretrain Data	Epoch	MIR
Vicuna-13B-v1.5	CLIP-L/336	MLP-2x	LCS-558K	1epoch	2.583
Vicuna-7B-v1.5	CLIP-L/336	MLP-2x	LCS-558K	1epoch	3.374
LLaMA-2-13B-Chat	CLIP-L/336	Linear	LCS-558K	1epoch	2.477
LLaMA-2-7B-Chat	CLIP-L/336	Linear	LCS-558K	1epoch	3.699

Table 5. MIR values of LVLMs that have different scale of LLMs.

model. The results in Table 2 proves that MoCa can also improves pre-training quality in 13B model.

References

[1] 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. 1

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 2

[4] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006. 2

[5] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024. 1

[6] 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*, 2023. 1

[7] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024. 1

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, 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*, pages 24185–24198, 2024. 1

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1

[10] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 1

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 2

[12] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade

Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 1

[13] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019. 2

[14] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 1

[15] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1

[16] 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, 2023. 1

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[18] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1

[19] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

[21] 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. 1

[22] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 1

[23] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 1

[24] Yuxuan Qiao, Haodong Duan, Xinyu Fang, Junming Yang, Lin Chen, Songyang Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Prism: A framework for decoupling and assessing the capabilities of vlms. *arXiv preprint arXiv:2406.14544*, 2024. 1

[25] 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. 1

[26] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1

[27] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 1

[28] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1

[29] Lai Wei, Zhiquan Tan, Chenghai Li, Jindong Wang, and Weiran Huang. Diff-erank: A novel rank-based metric for evaluating large language models. *arXiv preprint arXiv:2401.17139*, 2024. 1

[30] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 1

[31] 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. 1

[32] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Haodong Duan, Songyang Zhang, Shuangrui Ding, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023. 1

[33] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023. 1

[34] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1