

Data or Language Supervision: What Makes CLIP Better than DINO?

Yiming Liu^{1,2,*}, Yuhui Zhang^{1,*}, Dhruva Ghosh¹,

Ludwig Schmidt^{1,†}, Serena Yeung-Levy^{1,†}

¹Stanford University, ²Tsinghua University

*equal contribution, †equal advising

{ymingliu, yuhuiz, djghosh, ludwigsc, syyeung}@stanford.edu

Abstract

CLIP outperforms self-supervised models like DINO as vision encoders for vision-language models (VLMs), but it remains unclear whether this advantage stems from CLIP’s language supervision or its much larger training data. To disentangle these factors, we pre-train CLIP and DINO under controlled settings—using the same architecture, dataset, and training configuration—achieving similar ImageNet accuracy. Embedding analysis shows that CLIP captures high-level semantics (e.g., object categories, text), while DINO is more responsive to low-level features like colors and styles. When integrated into VLMs and evaluated on 20 VQA benchmarks, CLIP excels at text-intensive tasks, while DINO slightly outperforms on vision-centric ones. Variants of language supervision (e.g., sigmoid loss, pre-trained language encoders) yield limited gains. Our findings provide scientific insights into vision encoder design and its impact on VLM performance.¹

1 Introduction

Vision-language models, such as GPT-4o (OpenAI, 2023) and Claude (Anthropic, 2024), have demonstrated transformative capabilities in interpreting and reasoning over visual inputs. These models typically consist of a vision encoder, a language model, and a connector module bridging the two (Liu et al., 2023). The vision encoder—serving as the “eyes” of the system—plays a critical role in transmitting visual information to the language model, which acts as the “brain” that interprets it.

Recent studies have shown that CLIP (Radford et al., 2021), particularly its variant SigLIP (Zhai et al., 2023), has emerged as the most effective vision encoder for building VLMs, significantly outperforming DINO-based counterparts across various domains (Karamcheti et al., 2024; Tong et al.,

¹Code available at <https://github.com/leo1oel/Controlled-CLIP-DINO>.

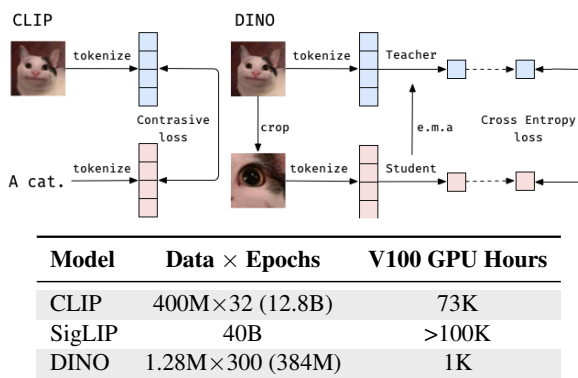


Figure 1: CLIP and DINO represent two predominant paradigms of vision encoders, differing in two key aspects: (1) CLIP is trained with language supervision, whereas DINO uses image-only self-supervision; (2) CLIP and its variant SigLIP are trained on datasets that are up to 100 times larger than those used for DINO. These differences make it difficult to disentangle whether CLIP’s superior performance in vision-language models stems from its training objective or the scale of its training data.

2024a). These two types of vision encoders represent two major paradigms: CLIP is trained using image-text contrastive learning, while DINO employs image-only self-supervised learning (Figure 1, top).

This observation raises a fundamental question: **Is CLIP’s superior performance primarily due to its language-supervised training objective, or is it simply a result of its significantly larger training dataset?** Despite the difference in supervision, CLIP models are often trained on datasets that are up to 100 times larger than those used for DINO (Figure 1, bottom). This substantial discrepancy in data scale makes it difficult to disentangle the effects of training objective from those of dataset size.

To isolate these factors, we conduct a controlled study by training CLIP and DINO vision encoders under identical conditions: the same architecture (ViT-B/16), dataset (a 10M-image subset of Data-

Models	General		Fine-grained			Robustness	
	ImageNet	CIFAR10	Stanford Cars	Flowers	CUB	ImageNetV2	CIFAR10.1
Official CLIP	79.5	93.4	80.8	89.7	74.9	68.9	87.3
Official DINO	76.1	93.5	61.2	83.2	71.0	65.5	84.7
Controlled CLIP	65.8	90.7	74.7	78.7	52.3	53.0	82.8
Controlled DINO	66.4	92.1	54.1	80.7	43.0	53.5	86.0

Table 1: **Linear probing accuracy (%) of controlled CLIP and DINO.** Trained under identical settings except for the presence of language supervision in CLIP, both models perform similarly on general and robustness benchmarks. However, CLIP shows significantly higher accuracy on fine-grained classification tasks, highlighting the benefit of language supervision in distinguishing visually similar categories.

Comp (Gadre et al., 2023)), and training configurations (20 epochs). Notably, the resulting models achieve comparable ImageNet (Deng et al., 2009) linear probing accuracy (CLIP: 65.8%, DINO: 66.4%), ensuring a fair basis for comparison.

Using these controlled encoders, we first investigate how language supervision alters the embedding space. We identify and analyze image pairs where CLIP and DINO produce significantly different similarity scores. Our analysis reveals that CLIP is more sensitive to high-level visual semantics—such as object type and embedded text—while DINO is more responsive to low-level visual attributes like colors and styles.

We then integrate these controlled encoders into the LLaVA (Liu et al., 2023) and train the resulting VLMs under identical settings. Evaluated on 20 VQA benchmarks, LLaVA-CLIP significantly outperforms LLaVA-DINO on text-intensive tasks (e.g., questions involving tables or charts), achieving a 7.5% performance gain. LLaVA-DINO performs slightly better on some visually grounded tasks but matches LLaVA-CLIP on most others.

To further probe the effect of language supervision, we explore two additional questions: (1) Do different supervision objectives, such as CLIP vs. SigLIP, lead to performance differences? (2) Does using a pre-trained language encoder during CLIP training yield a stronger vision encoder? In both cases, we find the answer to be no.

In summary, our study examines how vision encoders influence the performance of vision-language models. Through carefully controlled experiments, we uncover the representational differences induced by language supervision and their downstream effects, thereby offering the community deeper scientific insights into designing vision-centric vision-language systems.

2 Training Controlled CLIP and DINO

In this section, we describe how we train CLIP and DINO under controlled settings, ensuring that the only difference lies in the supervision signal. We then evaluate their performance on various image classification benchmarks.

Experimental setup. To isolate the effect of supervision, we align all other factors in CLIP and DINO training. 1) Architecture: We use ViT-B/16 as the backbone for both models; 2) Dataset: We train on a 10M subset of the DataComp (Gadre et al., 2023) image-caption dataset. All images are center-cropped and resized to 224×224 . For CLIP, we use the associated captions as language supervision; for DINO, no textual input is provided; 3) Training Configuration: Both models are trained from scratch for 20 epochs using the AdamW optimizer, a learning rate of $1e-3$, and cosine learning rate decay. Training is conducted on 4 A100 GPUs over 3 days.

Results. After training, we evaluate the encoders using linear probing on standard image classification benchmarks—a widely adopted approach for assessing vision encoder quality. As shown in Table 1, the models perform similarly on general classification tasks such as ImageNet (Deng et al., 2009) and CIFAR-10 (Krizhevsky, 2009). However, the difference becomes more pronounced on fine-grained classification benchmarks: CLIP significantly outperforms DINO on Stanford Cars (Krause et al., 2013) (74.7% vs. 54.1%, +20.6%) and CUB (Wah et al., 2011) (52.3% vs. 43.0%, +9.3%), despite being trained on the same image data. This suggests that language supervision is especially helpful for tasks requiring detailed semantic distinctions. For robustness evaluation, performance is comparable between CLIP and DINO, aligning with previous findings (Fang et al., 2022). Overall, these results indicate that while training

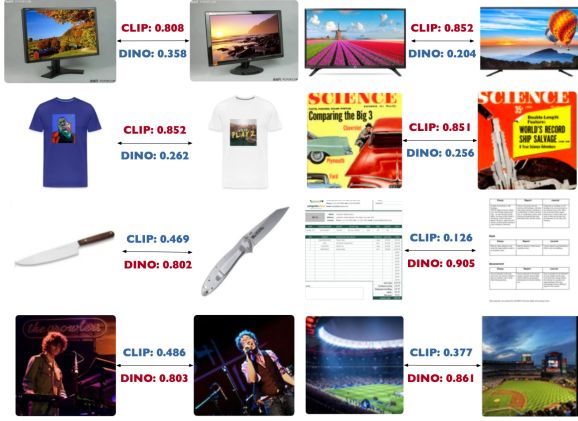


Figure 2: **Embedding analysis of CLIP and DINO.** The top two image pairs exhibit high cosine similarity according to CLIP but low similarity under DINO, suggesting that CLIP is more attuned to high-level semantics such as object categories and embedded text. In contrast, the bottom pairs show the opposite pattern, indicating that DINO is more sensitive to low-level features like object colors and visual styles.

data scale governs general classification and robustness, language supervision provides substantial benefits for fine-grained recognition tasks where subtle visual differences must be captured.

3 Embedding Analysis

To gain deeper insight into how language supervision shapes the embedding space, we conduct a fine-grained embedding analysis comparing CLIP and DINO. Unlike coarse-grained metrics like classification accuracy, this analysis reveals how each model organizes visual information.

Method. Similar to Tong et al. (Tong et al., 2024b), we analyze pairs of images in DataComp-10M where CLIP and DINO produce highly divergent similarity scores, revealing systematic differences in representation. Specifically, we identify two types of image pairs:

$$g_1 = (\text{clip_sim} > 0.8) \wedge (\text{dino_sim} < 0.5)$$

$$g_2 = (\text{dino_sim} > 0.8) \wedge (\text{clip_sim} < 0.5)$$

These selected pairs help isolate cases where the two models disagree in their embeddings.

Results. Figure 2 illustrates representative examples for our analysis. CLIP shows strong alignment with high-level semantic features such as object identity and textual content. It consistently groups images by object type or embedded texts, even across variations in visual style or

context—suggesting that language supervision enhances semantic abstraction. In contrast, DINO is more sensitive to low-level visual cues like color schemes, and is more invariant to orientation change. We provide a quantitative validation for these observations in Appendix B. These findings highlight that CLIP learns embeddings that are more semantically meaningful, while DINO emphasizes visual similarity, likely due to its self-supervised objective.

4 VLM Analysis

After training the controlled CLIP and DINO encoders, we incorporate them into the LLaVA-1.5 framework to investigate how vision encoder choice impacts the performance of VLMs.

Experimental Setup. We use LLaVA-1.5 with its vision encoder replaced by either controlled CLIP or DINO. Training consists of pretraining followed by visual instruction tuning. During training, we save checkpoints every 500 steps and evaluate each on VMCBench (Zhang et al., 2025)—a unified multiple-choice visual question answering benchmark composed of 20 datasets—to simplify evaluation. Since test set labels are not publicly available, we select the best checkpoint based on validation performance and report validation results. All training configurations are kept identical for both CLIP and DINO versions.

Results. Figure 3 presents performance across the 20 VMCBench subsets. **CLIP and DINO perform comparably on most tasks:** On general VQA and reasoning tasks, both encoders yield similar results. For instance, DINO achieves 41.5% accuracy on reasoning tasks versus CLIP’s 41.2%; for general VQA, CLIP slightly edges out DINO at 46.2% versus 46.0%. In document and chart understanding (Doc&Chart), performance is nearly identical: 33.2% for CLIP vs. 33.1% for DINO. These small differences suggest that both encoders are similarly effective in broad VLM tasks. **CLIP excels in text-intensive visual tasks:** The most notable difference appears in OCR-based benchmarks. On average, LLaVA-CLIP achieves 47.5% on OCRVQA (Mishra et al., 2019) and TextVQA (Singh et al., 2019), while LLaVA-DINO reaches only 40.0%, a substantial 7.5 percentage-point gap. This result indicates that language supervision in CLIP enhances its ability to extract and reason over textual content embedded in images—a key capa-

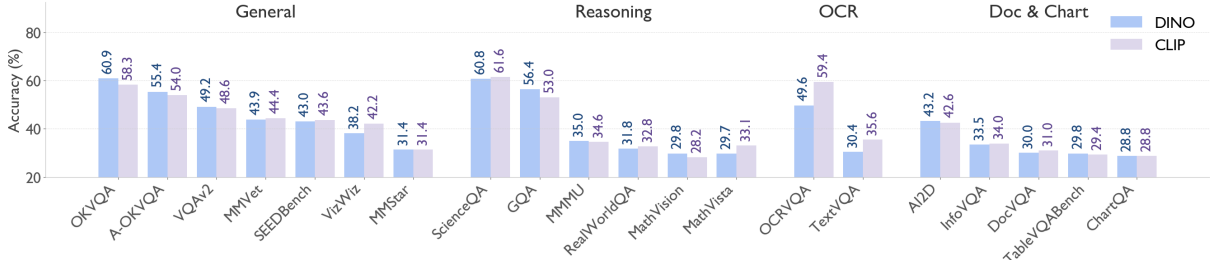


Figure 3: **VLM analysis of CLIP and DINO.** We integrate the controlled CLIP and DINO encoders into LLaVA-1.5 and evaluate on 20 subsets of the VMCBench benchmark. Results show that LLaVA-CLIP significantly outperforms LLaVA-DINO on OCR tasks by 7.5%, while their performance is largely comparable on other tasks.

CLIP	SigLIP Loss	Pretrained LM (Vicuna)
41.4	40.8	40.5

Table 2: Alternative language supervision objectives or using a pretrained text encoder do not improve CLIP performance when used in vision-language models.

bility for text-heavy visual understanding.

5 Exploring Better Language Supervision

Given that language supervision (1) improves fine-grained image classification, (2) encourages high-level semantic alignment, and (3) enhances OCR task performance in VLMs, we further explore whether alternative forms of language supervision can yield stronger vision encoders.

Experimental Setup. We explore two directions to improve CLIP’s language supervision. First, we replace the standard contrastive loss with the sigmoid-based SigLIP loss to examine whether the training objective affects performance. Second, we substitute the randomly initialized text encoder in CLIP with a frozen, pretrained Vicuna-7B (Zheng et al., 2023) model to assess the value of stronger language priors. After training, we integrate each encoder into the LLaVA-based VLM (as described previously) and evaluate on VMCBench.

Results. As shown in Table 2, neither modification outperforms the baseline CLIP model. Both the SigLIP loss and the pretrained Vicuna-based encoder yield slightly lower average accuracy. These results suggest that while language supervision is critical, the specific form—whether via objective function or pretrained language model—may offer limited additional benefit, consistent with recent observations in the literature (Huang et al., 2024).

6 Related Works

Vision-Language Models. Recent years have seen rapid advances in Vision-Language Models

(VLMs), with architectures such as LLaVA (Liu et al., 2023) and Qwen2.5-VL (Bai et al., 2025) demonstrating increasingly sophisticated multi-modal capabilities. These models typically pair a vision encoder with a large language model (LLM), enabling joint reasoning over visual and textual inputs. In this framework, the vision encoder plays a critical role by converting images into representations that can be projected and processed by the LLM. Our work focuses on this vision encoder component, aiming to understand how its training affects downstream VLM performance.

Visual Representation Learning. Visual representation learning mainly follows two paradigms: self-supervised and language-supervised learning. Self-supervised approaches, such as DINO (Caron et al., 2021) and SimCLR (Chen et al., 2020), learn representations by predicting relationships between augmented views of the same image. In contrast, language-supervised methods—exemplified by CLIP (Radford et al., 2021), EVA-CLIP (Sun et al., 2023), and SigLIP (Zhai et al., 2023)—leverage image-text pairs to align visual and linguistic representations. These two families of methods not only differ in supervision strategy but also in the scale of training data. In this work, we systematically ablate which factor—supervision type or data scale—drives performance gains.

Design Choices in Vision-Language Models.

Several studies have investigated how architectural components, data curation strategies, and training configurations affect VLM performance (Karamcheti et al., 2024; Laurençon et al., 2024; McKinzie et al., 2024). Across these works, CLIP and its variants (e.g., SigLIP) consistently emerge as the most effective vision encoders. However, such findings are typically based on pre-trained models, which differ in supervision objectives, data size, and training setups—making it difficult to isolate the source of performance differences. In contrast, our work

trains CLIP and DINO under controlled conditions to isolate the effect of language supervision on vision encoder quality.

7 Conclusion

This work conducts a controlled study to disentangle the effects of language supervision and data scale on vision encoder performance in VLMs, offering insights into vision encoder design and its role in effective VLMs.

Acknowledgments

S.Y. is a Chan Zuckerberg Biohub — San Francisco Investigator.

Limitations

While our study is carefully controlled, it is limited to a 10M-image subset. Scaling these comparisons to billion-image datasets is a crucial next step for fully understanding the interplay between supervision type and data magnitude. A concurrent work addressed this question by scaling DINO and CLIP to 7B parameters on 8B data (Fan et al., 2025). Additionally, exploring hybrid approaches that strategically combine self-supervised and language-supervised signals remains a promising direction for advancing vision encoder design.

References

- Anthropic. 2024. [Introducing the next generation of claude](#). Technical report, Anthropic. Accessed 2025-05-17.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv: 2502.13923*.
- Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning (ICML)*.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- David Fan, Shengbang Tong, Jiachen Zhu, Koustuv Sinha, Zhuang Liu, Xinlei Chen, Michael Rabbat, Nicolas Ballas, Yann LeCun, Amir Bar, and Saining Xie. 2025. [Scaling language-free visual representation learning](#). *arXiv preprint arXiv: 2504.01017*.
- Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. 2022. Data determines distributional robustness in contrastive language image pre-training (clip). *arXiv preprint arXiv: 2205.01397*.
- Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M. Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, and 15 others. 2023. Datacomp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Liang Hu, Qi Dai, Xiyang Dai, Dongdong Chen, Chong Luo, and 1 others. 2024. Llm2clip: Powerful language model unlock richer visual representation. *arXiv preprint arXiv:2411.04997*.
- Siddharth Karamcheti, Suraj Nair, A. Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. 2024. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *Proceedings of the International Conference on Machine Learning (ICML)*.
- Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. Collecting a large-scale dataset of fine-grained cars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images. Technical report, University of Toronto.
- Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruvi Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug

- Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, and 13 others. 2024. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv: 2403.09611*.
- Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *International Conference on Document Analysis and Recognition (ICDAR)*.
- OpenAI. 2023. [Gpt-4v\(ision\) system card](#). Technical report, OpenAI. Accessed 2025-05-17.
- 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. In *Proceedings of the International Conference on Machine Learning (ICML)*.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv: 2303.15389*.
- Shengbang Tong, Ellis L Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024a. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024b. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. Technical report, California Institute of Technology.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *arXiv preprint arXiv: 2407.10671*.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Alejandro Lozano, Ludwig Schmidt, and Serena Yeung-Levy. 2025. Automated generation of challenging multiple choice questions for vision language model evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, E. Xing, Haoteng Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Training Curves

We provide the training loss curves for both CLIP and DINO under the controlled setup. As shown in Figure 4, both models converge smoothly within 20 epochs, with no signs of overfitting or instability.

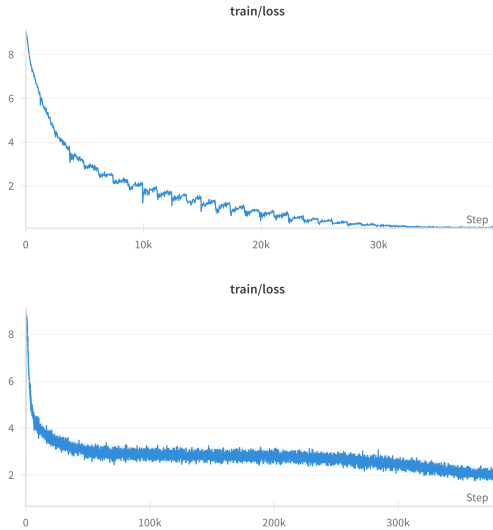


Figure 4: Training loss curves for the controlled CLIP (top) and DINO (bottom) models.

B Quantitative Validation of Embedding Space

To provide quantitative support for the claims made in our main embedding analysis (Section 3), we conducted two experiments measuring the cosine similarity within the controlled CLIP and DINO embedding spaces.

Experiment 1: Sensitivity to Semantic Content (Text). To test the models’ ability to distinguish between high-level semantic concepts, we created a small dataset of images where each image contained a unique alphabet letter or number (‘A’, ‘B’, ‘1’, etc.). We then computed the average pairwise cosine similarity between the embeddings of these semantically distinct images.

Results: The average similarity for DINO was 0.877, while for CLIP it was significantly lower at 0.713.

Conclusion: The lower similarity score for CLIP demonstrates that its representations for different semantic symbols are more separable and distinct. This quantitatively confirms that CLIP’s embedding space is more structured around the semantic identity of the visual content.

Experiment 2: Sensitivity to Visual Patterns.

To measure sensitivity to low-level features, we performed a similar analysis on a dataset of images containing simple, repeating visual patterns (e.g., grids, dots, checkers), where semantic content was minimal.

Results: In this case, the trend reversed. The average similarity for DINO was 0.478, while for CLIP it was 0.497.

Conclusion: The lower similarity score for DINO indicates that its representation space separates these low-level visual patterns more effectively. This provides quantitative support for our claim that DINO is more sensitive to visual structure.

Together, these quantitative results align perfectly with our qualitative analysis, providing a robust and comprehensive picture of how language supervision shapes visual representations compared to self-supervision.

C Using Qwen2-7B as the LLM Backbone

To further examine the interaction between vision encoders and language models, we evaluate the performance of our controlled CLIP and DINO encoders using Qwen2-7B (Yang et al., 2024) as the LLM backbone, in comparison to Vicuna-7B. Results are summarized in Table 3.

Improved General VQA Performance with Qwen2-7B. When paired with Qwen2-7B, CLIP demonstrates an advantage in general VQA tasks, achieving 57.90% accuracy compared to DINO’s 54.02%—a 3.88 percentage point gain. This contrasts with the Vicuna-7B setting, where CLIP and DINO achieved nearly identical results in the same category (46.23% vs. 46.20%). These results suggest that Qwen2-7B may better leverage CLIP’s high-level semantic representations for tasks requiring holistic scene understanding.

Model	General	Reason	Doc/Chart	OCR	Avg
CLIP + Vicuna	46.23	41.17	33.15	47.50	41.44
DINO + Vicuna	46.20	41.50	33.07	40.00	40.71
CLIP + Qwen2	57.90	47.74	40.62	51.40	49.69
DINO + Qwen2	54.02	47.56	39.86	47.59	47.72

Table 3: Performance on VMCBench using different vision encoder and LLM backbone combinations. Qwen2-7B leads to stronger performance across most categories, especially when paired with CLIP.