

A Training Data and Model Efficacy

Across both vision and text regression, we observed double descent, a phenomena which has been studied in depth in prior literature [Bach, 2023, Bartlett et al., 2020, Mei and Montanari, 2022, Schaeffer et al., 2024]. As seen in Figure 1, the task matrix performance rises with the number of images used in construction, but declines sharply near the full dimension of the embedding space (768 in CLIP). When the number of input samples is equal to the embedding dimension, a unique exact solution exists. In practice, we employed many more images than the embedding dimension, opting to use the full training dataset for task matrix construction.

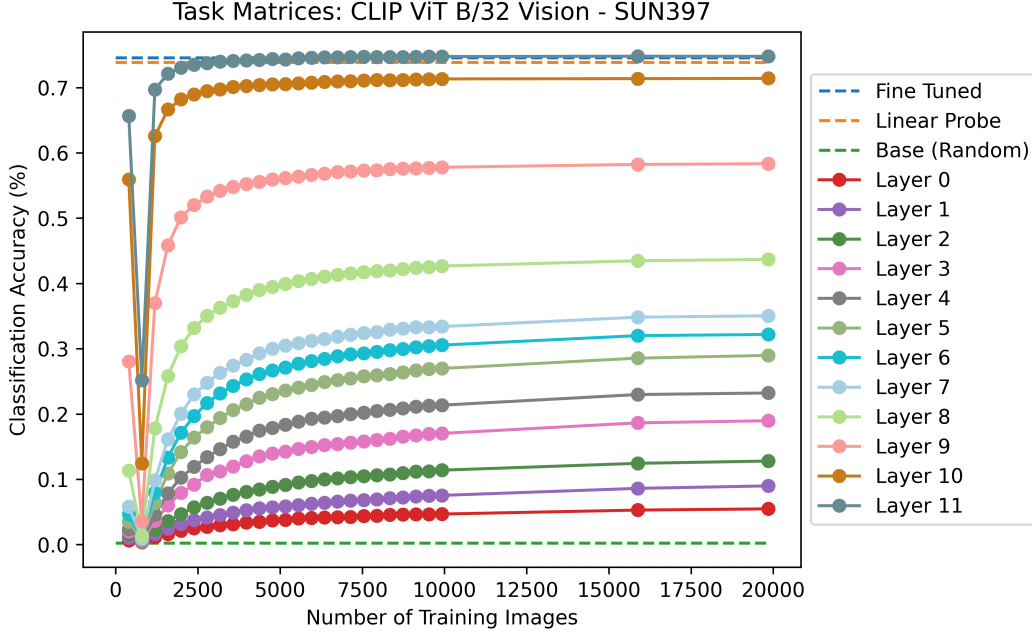


Figure 1: Classification Accuracy of SUN397 Across Layers and Training Images.

B CLIP ViT-B/32: Normalized Multi-Class Augmentation Results

We show multi-task performance in Figure 2. A single task matrix across multiple datasets remains highly effective over individual evaluations.

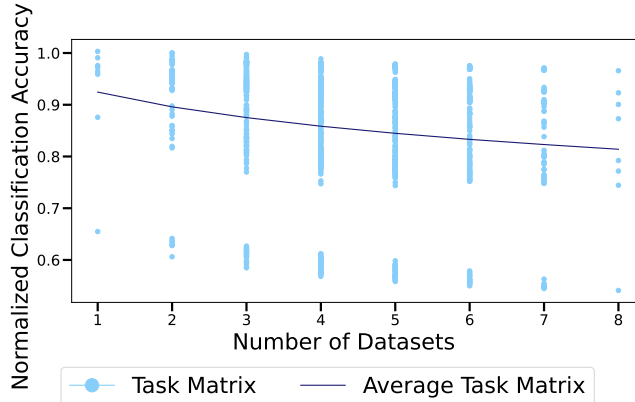


Figure 2: CLIP ViT Vision Full Train Multi-Task (1-8) Augmentation. Small drops in accuracy are seen (from 92%-81%). All results are from layer 11 and are over five trials.

C Ablation: Direct Readout from Base Model Results

Table 1: RoBERTa Base w/FT Classifier Ablation performance (%). The classifier head was taken from the fine-tuned model and used to directly read out from the base model. Task matrices for RoBERTa greatly exceed the results here, demonstrating the task matrix is necessary for the improved performance. RoBERTa (n=5, 95% CI). Layers are numbered 0-23.

Base w/ FT Classifier Method (classes)	Emotion (6)	HANS (2)	BLiMP (67)	Trec-6 (6)	SNLI (3)	ATIS (18)	Banking77 (77)
Full Train (best layer)	25.5±13.6 (23)	50.1±0.2 (23)	2.4±0.8 (23)	23.1±0.7 (23)	33.7±0.7 (23)	18.6±13.8 (23)	1.7±0.2 (23)

Table 2: Vision Base w/FT Classifier Ablation performance (%). The classifier head was taken from the fine-tuned model and used to directly read out from the base model. Task Matrix performance for CLIP exceed results here in all datasets, demonstrating that the task matrix is significant. CLIP ViT-B/32 vision tower (n=5, 95% CI). Layers are numbered 0-11.

Base w/ FT Classifier Method (classes)	DTD (47)	EuroSAT (10)	GTSRB (43)	MNIST (10)	RESISC (45)	Stanford Cars (196)	SUN397 (397)	SVHN (10)
Full Train (best layer)	59.4±1.4 (11)	65.5±5.4 (10,11)	45.5±5.7 (11)	48.9±1.5 (11)	73.6±2.4 (11)	53.8±1.8 (11)	65.3±1.1 (11)	24.3±2.2 (11)

D Ablation: Frozen Decoder Head

To validate that our approach does not rely on adapted components, we modify our technique to operate on a frozen decoder head, which means our technique does not require any finetuned model components. In Table 3, we show that vision results with a frozen decoder head closely match earlier performances in CLIP’s general results section. This establishes that employing a task matrix alone is **sufficient** for good performance.

Table 3: Frozen Decoder Head performance (%). The classifier head was randomly initialized and frozen while fine-tuning. CLIP ViT-B/32 vision tower (n=5). Layers are numbered 0-11.

Method (num. classes)	DTD (47)	EuroSAT (10)	GTSRB (43)	MNIST (10)	RESISC45 (45)	Cars (196)	SUN397 (397)	SVHN (10)
Task Matrix (best layer)	74.9±0.4 (11)	96.6±0.3 (6,7,9)	86.9±0.4 (11)	99±0.08 (7)	90.3±0.5 (11)	72±0.5 (11)	70.1±0.4 (11)	67.7±0.8 (8)
Fine-Tuned	75.9±0.7	98.4±0.6	99.01±0.1	99.4±0.06	94±0.7	78.6±1.2	66.5±0.1	96.3±0.1

E all-MiniLM-L12-v2: Sentence Transformer Results

Table 4: Task Matrix against text baselines (%), all-MiniLM-L12-v2 (n=5, 95% CI). Layers are zero-indexed

Method	Emotion	HANS	BLiMP	Trec-6	SNLI	ATIS	Banking77
(classes)	(6)	(2)	(67)	(6)	(3)	(18)	(77)
Base w/ FT Classifier	24.5±13.0	50.4±0.7	2.6±0.8	22.5±1.5	33.6±0.3	21.6±36.3	23.2±4.4
Linear Probe	66.8±0.6	76.0±0.8	38.1±1.7	75.1±1.3	56.7±0.3	94.9±0.6	90.9±0.9
Task Matrix (best layer)	63.5±1.0 (L11)	82.3±2.0 (L8)	50.0±5.0 (L3,4)	84.7±0.9 (L0,5)	64.5±0.2 (L7)	91.9±0.7 (L5)	88.3±1.2 (L9,10)
Fine-Tuned	81.7±1.4	99.4±1.2	60.5±3.6	93.2±0.5	85.1±0.1	93.5±0.4	89.0±1.0

F Task Matrices in data-scarce settings

We next investigate settings in which the majority of data is held out for both probes and the task matrix. Concretely, we finetune the model on a 20% split of the training data, and subsequently construct a task matrix with the same quantity of restricted data and a linear probe with the same quantity of restricted data. We find that task matrices are far more robust to changes in data quantity than linear probes, exhibiting a 82% improvement on ATIS and 81% improvement on Trec-6. We show similar results for CLIP.

F.1 RoBERTa: Data-Scarce Results

Table 5: Task Matrix against text baselines (%) with training samples limited to 20% of the original dataset. The results exhibit minimal relative differences from the full training results in Roberta’s general results section. RoBERTa (n=5, SNLI n=2, 95% CI). Layers are numbered 0-23.

Method	Emotion	HANS	BLiMP	Trec-6	SNLI	ATIS	Banking77
(classes)	(6)	(2)	(67)	(6)	(3)	(18)	(77)
Linear Probe	25.8±2.8	81.5±2.3	26.6±10.1	30.3±6.8	59.1	42.7±2.4	14.1±1.5
Task Matrix (best layer)	36.3±2.1 (1,2)	95.8±0.4 (16)	55.2±1.6 (4,5,6)	55.0±3.2 (11)	76.4 (18,19)	77.7±1.9 (4,5,6)	46.5±1.1 (3,4)
Fine-Tuned	63.9±1.6	100.0±0.0	70.2±1.9	76.7±5.0	87.1	91.9±1.6	79.0±1.1

F.2 CLIP ViT-B/32: Data-Scarce Results

Table 6: Task matrix performance against vision baselines (%) with training samples limited to 20% of the original dataset. The results here exhibit minimal differences from the full training results. CLIP-ViT-B/32 vision tower (n=5, 95% CI). Layers are zero-indexed.

Method	DTD	EuroSAT	GTSRB	MNIST	RESISC	Stanford Cars	SUN397	SVHN
(classes)	(47)	(10)	(43)	(10)	(45)	(196)	(397)	(10)
Linear Probe	67.2±1	94.6±0.3	84.3±0.4	98.2±0.1	88.2±0.3	62.8±0.5	66.8±0.3	62.5±0.6
Task Matrix (best layer)	61.9±0.7 (11)	95.5±0.7 (6,7,9)	85.8±0.3 (11)	98.9±0.1 (7,8)	89.7±0.3 (11)	50.9±1 (11)	68.1±0.2 (11)	64.5±0.8 (8)
Fine-Tuned	67.5±1	97.5±0.4	97.8±0.2	99.3±0.1	91.5±0.5	58.2±1	67±0.3	93.6±0.3

G DeiT: Comprehensive Vision Results

Table 7: Task matrix against vision baselines (%), DeiT-tiny-patch16-224 (n=5, 95% CI). Layers are zero-indexed.

Method	DTD	EuroSAT	GTSRB	MNIST	RESISC	Stanford Cars	SUN397	SVHN
(classes)	(47)	(10)	(43)	(10)	(45)	(196)	(397)	(10)
Base w/ FT Classifier	54.8±1.6	70.4±3	30±5.5	47.1±9.8	57.7±1.7	5.6±1	41.8±0.8	23.8±3
Linear Probe	63.3±0.3	93.4±0.05	66±0.2	95.6±0.08	79.1±0.2	26.3±0.1	49.3±0.3	46.5±0.5
Task Matrix (best layer)	62.5±0.6 (L10,11)	95.1±0.2 (L5)	64.9±1.1 (L7)	95.7±0.4 (L4,5)	77.9±0.3 (L9)	31.4±0.4 (L9)	48.9±0.1 (L11)	49.6±1 (L7)
Fine-Tuned	67.4±0.6	98±0.2	96.7±0.5	99.2±0.1	89.3±0.2	50.7±1.1	55.1±0.3	95.1±0.2

Table 8: Constrained-data task matrices against baselines (%). With training samples limited to 20% of the original quantity, results remain consistent. DeiT-tiny-patch16-224 (n=5, 95% CI). Layers are zero-indexed.

Method	DTD	EuroSAT	GTSRB	MNIST	RESISC	Stanford Cars	SUN397	SVHN
(classes)	(47)	(10)	(43)	(10)	(45)	(196)	(397)	(10)
Base w/ FT Classifier	42.6±1.2	71.4±1.7	34.6±3.9	43±10.4	55.9±1.8	4.4±0.6	31.7±0.5	22.7±3.5
Linear Probe	52.1±1.5	91.6±0.1	62.8±0.4	95.1±0.1	73.1±0.3	12.4±0.2	38.7±0.5	45.1±0.4
Task Matrix (best layer)	50.1±1.1 (L9,10,11)	94.1±0.5 (L5)	64.6±1 (L7)	95.7±0.1 (L4)	74±0.6 (L7,8,9)	11.8±0.9 (L9)	37.6±0.3 (L11)	50±1.4 (L4,7)
Fine-Tuned	50±0.6	95.5±0.8	91.8±1.2	98.7±0.09	80.4±0.8	12±1.3	40.5±0.4	91.3±0.5

Table 9: Task matrix performance against vision baselines (%). The classifier head was randomly initialized, frozen while fine-tuning, and used for evaluations. Results remain consistent and approach finetuned levels over all datasets. DeiT-tiny-patch16-224 (n=5, 95% CI). Layers are zero-indexed.

Method	DTD	EuroSAT	GTSRB	MNIST	RESISC	Stanford Cars	SUN397	SVHN
(classes)	(47)	(10)	(43)	(10)	(45)	(196)	(397)	(10)
Base w/ FT Classifier	3.1±0.6	14.5±3.9	4.2±1.9	12.8±3.3	2.9±0.3	0.6±0.1	0.3±0.1	11.7±2.5
Task Matrix (best layer)	59.7±0.5 (L8,9)	94.6±0.3 (L5)	63±1.4 (L7)	95.1±0.4 (L3,4,7)	75.3±1.2 (L7,8)	16±1.9 (L9)	24.7±1.6 (L10)	48.8±1.2 (L7)
Fine-Tuned	63±1.5	98.5±0.2	98.6±0.1	99.5±0.06	90.5±1	26.9±5	38±1	96.1±0.2

Table 10: Task matrix performance against vision baselines (%). The classifier head was randomly initialized, frozen while fine-tuning, and used for evaluations. With training samples limited to 20% of the original dataset, results remain consistent and approach or exceed finetuned levels over all datasets. DeiT-tiny-patch16-224 (n=5, 95% CI). Layers are zero-indexed.

Method	DTD	EuroSAT	GTSRB	MNIST	RESISC	Stanford Cars	SUN397	SVHN
(classes)	(47)	(10)	(43)	(10)	(45)	(196)	(397)	(10)
Base w/ FT Classifier	2.7±0.5	12±3	4.8±0.8	12.8±1.3	3.4±0.9	0.6±0.09	0.3±0.1	15±3
Task Matrix (best layer)	47.2±1.3 (L9,10)	94±0.1 (L5)	59.2±1.1 (L7)	94.9±0.4 (L4)	72.8±0.3 (L7)	0.8±0.1 (L3,8,9,11)	12±1 (L10)	50.6±1.1 (L7)
Fine-Tuned	30±0.8	95.8±0.7	93.9±0.9	98.8±0.1	76.9±1.3	0.5±0.1	2.6±0.4	91.1±0.7

H DINOv3 ViT-B/16: Single Task Results

We extended the datasets used in order to comprehensively evaluate a novel vision transformer. Specifically, we experiment on INaturalist-Mini 2021 Horn et al. [2021], Cifar10 Krizhevsky and Hinton [2009], Cifar100 Krizhevsky and Hinton [2009], and Food101 Bossard et al. [2014]. The task matrix generally exhibits lower results than linear probes. We posit the task matrix’s lower performance to DINOv3 ViT-B/16’s strong pre-trained backbone, and the lack of middle-layer enrichment in vision models.

Table 11: Task Matrix against vision baselines (%), DINOv3 ViT-B/16 (n=5, 95% CI). Layers are zero-indexed.

Method	DTD	EuroSAT	GTSRB	MNIST	RESISC	Stanford Cars	SUN397	SVHN
(classes)	(47)	(10)	(43)	(10)	(45)	(196)	(397)	(10)
Linear Probe	83.5±0.1	97±0.09	85.7±0.2	98.7±0.03	92.7±0.1	93.8±0.1	77.2±0.09	66.9±0.1
Task Matrix (best layer)	82.9±0.3 (11)	97.1±0.1 (6,9,11)	86.3±0.7 (11)	98.9±0.1 (8)	91.4±0.1 (11)	93.7±0.09 (11)	76.7±0.4 (11)	63.1±1 (8)
Fine-Tuned	84.5±0.2	98.8±0.1	98.9±0.1	99.5±0.1	95.7±0.2	94.4±0.09	78.1±0.3	97±0.1

Table 12: Task Matrix against vision baselines (%), DINOv3 ViT-B/16 (n=5, 95% CI). Layers are zero-indexed.

Method	INaturalist-2021	Cifar10	Cifar100	Food101
(classes)	(10000)	(10)	(100)	(101)
Linear Probe	60.9±0.7	98±0.01	88.5±0.1	93.3±0.08
Task Matrix (best layer)	59.1±0.6 (11)	97.7±0.1 (11)	86.7±0.5 (11)	92.8±0.2 (11)
Fine-Tuned	68.4±0.9	98.9±0.1	92±0.5	93.4±0.2

I Text Dataset descriptions

Emotion [Saravia et al., 2018]: A text classification dataset containing English Twitter messages labeled with six basic emotions (anger, fear, joy, love, sadness, and surprise), designed to evaluate models’ ability to recognize emotional content in social media text.

HANS [McCoy et al., 2019]: A diagnostic dataset for natural language inference that systematically tests syntactic heuristics by containing examples where lexical overlap, subsequence, and constituent heuristics fail, revealing models’ reliance on spurious statistical patterns rather than genuine linguistic understanding.

BLiMP [Warstadt et al., 2020]: The Benchmark of Linguistic Minimal Pairs for English, consisting of 67 sub-datasets with 1,000 minimal pairs each that isolate specific contrasts in syntax, morphology, or semantics, enabling targeted evaluation of models’ grammatical knowledge. To extend BLiMP for classification, we treated minimal pairs from the 67 grammatical phenomena categories as sentence-level classification problems.

TREC-6 [Li and Roth, 2002]: A question classification dataset containing 5,500 labeled questions divided into 6 coarse semantic categories (abbreviation, entity, description, human, location, numeric) for open-domain, fact-based question answering systems.

SNLI [Bowman et al., 2015]: The Stanford Natural Language Inference corpus containing 570k human-written English sentence pairs manually labeled for entailment, contradiction, and neutral relationships, serving as a foundational benchmark for natural language understanding.

ATIS [Hemphill et al., 1990]: The Airline Travel Information Systems dataset consisting of audio recordings and transcripts of humans asking for flight information, containing 17 unique intent categories for evaluating spoken language understanding systems.

Banking-77 [Casanueva et al., 2020]: A fine-grained intent detection dataset in the banking domain comprising 13,083 customer service queries labeled with 77 distinct intents, designed to evaluate models’ ability to understand specific user intentions in specialized domains.

J Vision Dataset Handling

For the standard 8 classification datasets used across CLIP ViT B/32 Vision Tower, DeiT-tiny-patch16-224, and DINOv3 ViT-B/16, we utilized the publicly available dataset, "The Eight Image Classification Tasks" (Tangake).

DTD, MNIST, Stanford Cars, and SVHN contain the full number of original dataset images, while SUN397 is the 50-class split for both training and testing partitions. EuroSAT, GTSRB, and RE-SISC45 contain 2,700, 12,569, and 6,300 fewer total images than the full original dataset respectively.

For DINOv3 ViT-B/16, we utilized the full datasets for INaturalist 2021, Cifar10, Cifar100, and Food101.

References

- Francis Bach. High-dimensional analysis of double descent for linear regression with random projections, 2023. URL <https://arxiv.org/abs/2303.01372>.
- Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020. doi: 10.1073/pnas.1907378117.
- Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. Association for Computational Linguistics, 2015.
- Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on NLP for ConvAI - ACL 2020*, 2020.
- Charles T. Hemphill, John J. Godfrey, and George R. Doddington. The ATIS spoken language systems pilot corpus. In *Proceedings of the workshop on Speech and Natural Language*, pages 96–101, 1990.
- Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisín Mac Aodha. Benchmarking representation learning for natural world image collections, 2021. URL <https://arxiv.org/abs/2103.16483>.

- Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL <http://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>.
- Xin Li and Dan Roth. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.
- R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448. Association for Computational Linguistics, 2019.
- Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022. doi: 10.1002/cpa.22008.
- Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697. Association for Computational Linguistics, 2018.
- Rylan Schaeffer, Zachary Robertson, Akhilan Boopathy, Mikail Khona, Kateryna Pistunova, Jason W. Rocks, Ila R. Fiete, Andrey Gromov, and Sanmi Koyejo. Double descent demystified: Identifying, interpreting & ablating the sources of a deep learning puzzle. ICLR Blogposts, May 2024. <https://iclr-blogposts.github.io/2024/blog/double-descent-demystified/>.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020.