Probing Relational Reasoning in Pre-Trained ViTs

Madeleine Fenner, Jahan Khan, Erica Shivers, Carlos Freire

INTRODUCTION

Large pre-trained foundation models such as DINOv3 [1] and CLIP [2] provide generalized visual embeddings that have been adopted for a variety of downstream tasks.

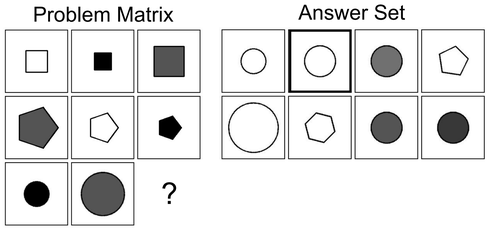
DINOv3 is a vision model with a vision transformer (ViT) backbone, while CLIP is a multimodal model for learning joint vision and language features, usually employing a ViT image encoder. Both models have been shown to achieve state-of-the-art (SOTA) performance in competitive vision tasks, indicating their image embeddings contain rich visual features. Exactly what nuanced features these image embeddings contain, and at what model layers they emerge, is less understood [3].

We investigate whether embeddings from DINOv3 and CLIP capture visual relationships within the image using the RAVEN-10000 dataset with a custom rule classification task. We probe different layers of the models to evaluate where relational context emerges. These results are important for informing suitable use cases of embeddings.

DATASET

Used RAVEN-10000, a dataset for evaluating relational and analogical visual reasoning [4].

RAVEN is made up of Raven's Progressive Matrices (RPMs), puzzles where each row follows a set of relational rules, and the image which completes the set must be identified.



Rules
Type (Shape)
- **Constant**
- Progression
- Distribute 3

Size
- Constant
- Progression
- Arithmetic
- **Distribute 3**

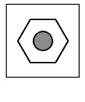
Color
- Constant
- Progression
- Arithmetic
- **Distribute 3**

RAVEN has 7 problem configurations, each with 10,000 RPMs. This study focused on 2 of the configurations:

Center

A single shape in the center of the image
Rules: Type, Size, Color

Out-In Center

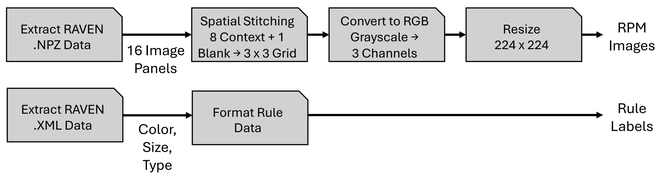


A single shape in the center of another shape
Rules: Outer - Type, Size
Inner - Type, Size, Color

METHODS

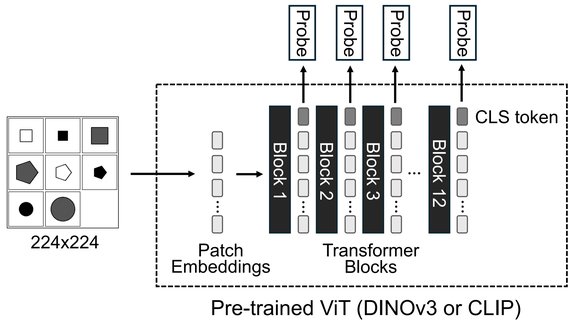
A core challenge in utilizing the RAVEN dataset with pre-trained vision transformers is their architectural constraint of processing single images. To evaluate relational reasoning in a pre-trained context, we defined a new rule prediction task: Given an RPM puzzle, classify the Type, Size, and Color rules which govern the arrangement.

Preprocessing



To preprocess the dataset, we stitched together the 8 context images for each RPM into a single 3x3 grid image and formatted the rule data into labels.

Probe Experiments



Models

DINOv3 - dinov3-vits16-pretrain-lvd1689m
CLIP (Vision Model Only) - clip-vit-base-patch32

Probes

Linear	Linear layer
MLP	Linear layer (input → 256)
	ReLU
	Linear layer (256 → output)

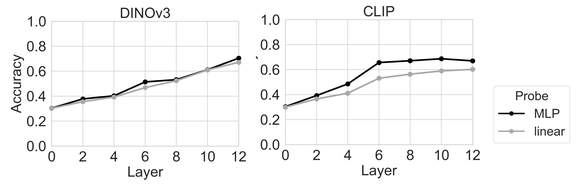
Hyperparameters

Learning rate	.001
Batch size	32
Epochs	45
Optimizer	Adam
Loss function	CrossEntropy

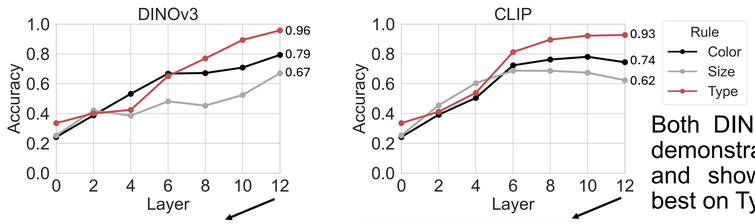
We employed a layer-wise probing methodology on two SOTA vision transformers: CLIP and DINOv3. We preprocessed the complete image set per configuration and divided images into 6,000 for training and 2,000 for testing, then extracted the CLS token embeddings for each layer of both base models. We trained a linear and MLP probe on the extracted embeddings to predict the underlying rules for each puzzle. This approach allowed us to investigate each layer's accessibility to color, shape, and size information at various depths of the network.

RESULTS

When comparing average accuracy between models over all tasks, using a Linear versus MLP probe impacted CLIP accuracy more than DINO accuracy. The following results use the MLP probe:



Center Configuration



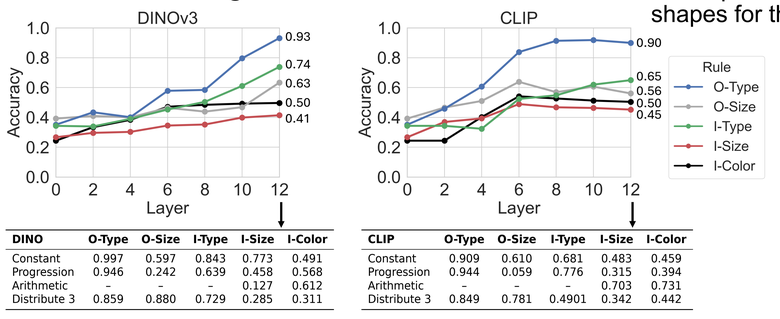
DINO	Type	Size	Color
Constant	0.978	0.735	0.856
Progression	0.982	0.609	0.782
Arithmetic	–	0.733	0.825
Distribute 3	0.914	0.605	0.701

CLIP	Type	Size	Color
Constant	0.967	0.563	0.791
Progression	0.889	0.782	0.563
Arithmetic	–	0.679	0.746
Distribute 3	0.929	0.459	0.865

Both DINO and CLIP embeddings demonstrate relational reasoning and show similar patterns, doing best on Type, then Color, then Size.

CLIP relational features appear at earlier layers, but accuracy can slightly decrease by the last layer.

Out-In Center Configuration



DINO	O-Type	O-Size	I-Type	I-Size	I-Color
Constant	0.997	0.597	0.843	0.773	0.491
Progression	0.946	0.242	0.639	0.458	0.568
Arithmetic	–	–	–	0.127	0.612
Distribute 3	0.859	0.880	0.729	0.285	0.311

CLIP	O-Type	O-Size	I-Type	I-Size	I-Color
Constant	0.909	0.610	0.681	0.483	0.459
Progression	0.944	0.059	0.776	0.315	0.394
Arithmetic	–	–	–	0.703	0.731
Distribute 3	0.849	0.781	0.4901	0.342	0.442

Both performed worse on inner shapes for the Out-In configuration.

DISCUSSION

Our study found that both CLIP and DINO embeddings contain relational information around types of shapes, size, and color to varying degrees. Rules relating to size had the lowest prediction accuracy for both models, suggesting the embeddings may not be suitable for tasks that depend on precise size comparisons. DINO generally outperformed CLIP, indicating semantic alignment of CLIP embeddings did not offer any advantage.

Limitations include that the RAVEN dataset uses synthetic images rather than natural images, which would align more closely with what the models trained on and will encounter for natural vision tasks. Future work could extend this implementation, trying other RAVEN configurations, or variations on the task such as using single rows. Using a reasoning head, probing could be expanded to full RPM puzzle reasoning where the target answer is predicted. Also, more difficult datasets such as DeepMind's Progressive Reasoning Matrices could be tried.

References
[1] O. Siméoni et al., "DINOv3," Aug. 13, 2025, arXiv: arXiv:2508.10104. doi: 10.48550/arXiv.2508.10104.
[2] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," Feb. 26, 2021, *arXiv*: arXiv:2103.00020. doi: 10.48550/arXiv.2103.00020.
[3] Oikarinen, T., & Weng, T.-W.. "CLIP-Dissect: Automatic Description of Neuron Representations in Deep Vision Networks". 2023. *arXiv [Cs.CV]*, 2023, arxiv.org/abs/2204.10965. arXiv.
[4] C. Zhang, F. Gao, B. Jia, Y. Zhu, and S.-C. Zhu, "RAVEN: A Dataset for Relational and Analogical Visual rEasoNing," Mar. 07, 2019, *arXiv*: arXiv:1903.02741. doi: 10.48550/arXiv.1903.02741.