DINO-QPM: Adapting Visual Foundation Models for Globally Interpretable Image Classification

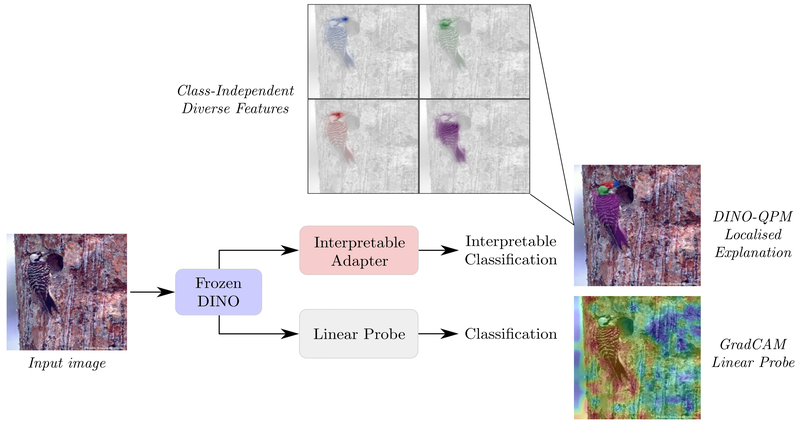
Robert Zimmermann Thomas Norrenbrock Bodo Rosenhahn

Leibniz University Hannover

CVPR JUNE 3-7, 2026 DENVER COLORADO

Introduction

- Visual Foundation Models (e.g. **DINOv2**) offer state-of-the-art feature extraction, with downstream classification typically done via a *linear probe*.
- **The problem:** safety-critical domains demand high interpretability, which standard linear probes lack.
- **Our solution:** we apply the Quadratic Programming Enhanced Model (QPM)[2] to build a lightweight adapter that transforms entangled DINOv2 representations into globally interpretable, contrastive features.

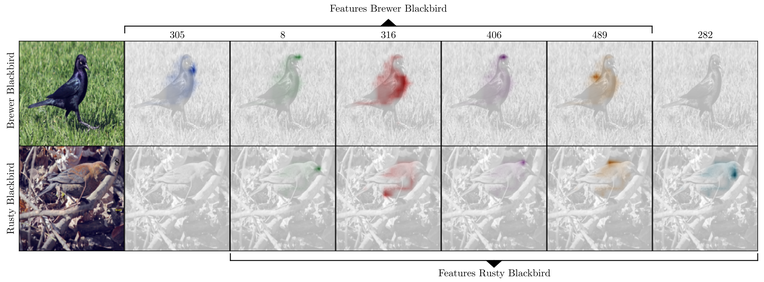


DINO-QPM compared to a standard linear probe: built on top of *strictly frozen* DINOv2 features, our lightweight adapter converts entangled patch embeddings into contrastive, diverse features.

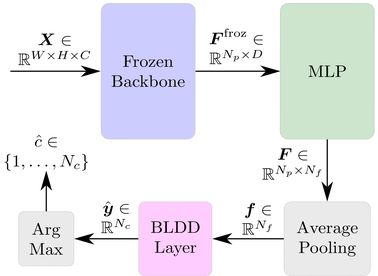
We achieve this through a sparse binary feature–class assignment that forms meaningful, class-shared concepts an unstructured classifier head cannot expose.

By replacing the conventional `CLS` token, on which the linear probe relies, with average-pooling over patch embeddings, these globally interpretable features can be *spatially localised* back to the input image.

Method



What it means to be **globally interpretable**: each class is represented by a small set of concepts (DINO-QPM's meaningful features, shown as saliency maps; here 5 per class), letting us compare classes based on their concepts. Visually similar classes can share many concepts; at least one always differs, which is what tells them apart.

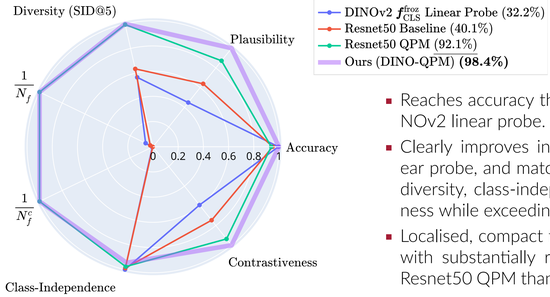


An input image $\boldsymbol{X}$ is encoded by a **frozen DINOv2 backbone** into per-patch embeddings $\boldsymbol{F}^{\text{froz}}$, which a lightweight **MLP** projects into task-specific feature maps $\boldsymbol{F}$. **Average pooling** reduces these to a feature vector $\boldsymbol{f}$. The **BLDD layer** has two states. During *dense training* it is a standard linear layer $\boldsymbol{W}\boldsymbol{f}$. During *fine-tuning* a binary selection vector $\boldsymbol{s}$ picks a subset of features from $\boldsymbol{f}$, and a sparse, binary projection $\boldsymbol{W}^{\text{sparse}}$ assigns exactly N_f^c of them to each class. Both $\boldsymbol{s}$ and $\boldsymbol{W}^{\text{sparse}}$ are found simultaneously by applying the QP from Norrenbrock et al. [2].

Three-stage Training: (1) Train the dense model with an unconstrained linear mapping $\boldsymbol{W} \in \mathbb{R}^{N_f\times N_c}$. (2) The QP [2] jointly derives the feature selection $\boldsymbol{s} \in \{0,1\}^{N_f}$ and the sparse binary mapping $\boldsymbol{W}^{\text{sparse}} \in \{0,1\}^{N_f\times N_c}$, assigning N_f^c features to each class. (3) Fine-tune the model with $\boldsymbol{W}^{\text{sparse}}$ fixed.

Results

All metrics evaluated on CUB-2011 [3]; Stanford Cars [1] results are in the paper.



- Reaches accuracy that slightly exceeds the DINOv2 linear probe.
- Clearly improves interpretability over the linear probe, and matches Resnet50 QPM [2] on diversity, class-independence and contrastiveness while exceeding it in plausibility.
- Localised, compact features ($N_f^c = 5$ per class), with substantially reduced training cost over Resnet50 QPM thanks to the frozen backbone.

Metrics

- **Accuracy.** Top-1 classification accuracy on the test set.
- **$1/N_f$, $1/N_f^c$.** Compactness: fewer features total and per class (dataset-independent).
- **Class-Independence.** Features generalise across classes rather than memorising single ones.
- **Diversity (SID@5).** Feature maps activate at distinct spatial positions.
- **Plausibility.** Feature map activations concentrate within the object's ground-truth region.
- **Contrastiveness.** Features are either on or off, forming a well-separated bimodal activation distribution over samples.

Conclusion

- **Lightweight interpretability adapter** on top of frozen visual foundation models (e.g. DINOv2), with no backbone fine-tuning required.
- **Localisable explanations** via average-pooling over patch tokens, quantified by our introduced Plausibility metric.
- **State-of-the-art interpretability** on par with Resnet50 QPM [2], at accuracy slightly exceeding the DINOv2 linear probe.

Code on GitHub

References

[1] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops*, 2013.

[2] T. Norrenbrock, T. Kaiser, S. Biswas, R. Manuvinakurike, and B. Rosenhahn. QPM: Discrete Optimization for Globally Interpretable Image Classification. In *The Thirteenth International Conference on Learning Representations*, 2025.

[3] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Contact: Robert Zimmermann, Leibniz University Hannover · `zimmerro@tnt.uni-hannover.de` · `rszimmermann.de`