Indian Institute of Technology Delhi

Department of Computer Science

COL-828
Advanced Computer Vision

Test Time Registers in Medical Imaging

Spandan Kukade - 2022CS51138
Yash Bansal - 2022CS51133

November 24, 2025

Contents

Abstract

Vision Transformers (ViTs) trained with self-supervised learning exhibit artifact tokens in later layers that accumulate disproportionately high attention and activation norms. While recent work has debated whether explicit register tokens improve model performance, the behaviour of naturally emerged registers in pretrained models remains underexplored. We systematically investigate register token detection and removal across CLIP and DINOv2 architectures on three diverse datasets: OrthoNet (orthopaedic implants), Pacemaker (device classification), and APTOS (diabetic retinopathy). Our experiments reveal that both architectures contain detectable register-like tokens, with DINOv2 exhibiting more severe artifacts (norms exceeding 600) than CLIP (norms around 150). Register removal consistently produces cleaner attention patterns with 83-96% reductions in artifact severity, yet performance impacts vary significantly by task and model, ranging from +4.44% to -1.33% accuracy. Medical imaging tasks requiring fine-grained discrimination show mixed benefits, while object recognition results depend on baseline artifact severity. These findings demonstrate that register tokens serve context-dependent functional roles rather than being universally beneficial or harmful, providing practitioners with empirical guidance for architecture decisions in vision transformer deployment.

1 Introduction

1.1 Problem Overview

Vision Transformers (ViTs) have become a dominant architecture in computer vision by adapting the transformer mechanism to process images as sequences of patches. While they achieve state-of-the-art performance on various tasks, recent work has identified an unexpected behaviour: the emergence of "artifact tokens" in the attention maps of trained models. These tokens accumulate high attention weights but don't correspond to meaningful image content, acting as information sinks during inference.

Darcet et al. [1] proposed adding explicit "register tokens" to provide designated locations for this global information aggregation, improving both performance and interpretability. However, Walmer et al. [2] later argued that registers may emerge naturally and that explicit training might be unnecessary. This contradiction, combined with limited testing across diverse domains, leaves open questions about when and whether register tokens are actually beneficial.

Understanding register token behavior has practical importance: removing unnecessary tokens could improve efficiency, while in interpretability-critical applications like medical imaging, we need to know whether registers help or hurt attention pattern clarity. Our work investigates these questions empirically across multiple datasets and model architectures.

1.2 Research Questions

We address the following questions:

1. **Register Token Detection:** Do pre-trained Vision Transformer models contain detectable register-like tokens? How does this differ between pre-trained and fine-tuned models across CLIP and DINO architectures?

2. **Performance Impact:** Does removing register tokens from pre-trained models improve or degrade classification accuracy? Is this effect consistent across architectures and domains?

3. **Attention Patterns:** How do register tokens affect attention mechanisms? Do they improve interpretability by localising artifacts or do they fragment attention in problematic ways?

4. **Domain Dependence:** Do register tokens behave differently in general classification versus specialised medical imaging tasks that require fine-grained feature discrimination?

We evaluate these questions across three datasets: Pacemaker (device classification), OrthoNet (orthopaedic implants), and APTOS (diabetic retinopathy detection). We examine both CLIP and DINO architectures in their pre-trained and fine-tuned states. Our experiments detect register tokens in all model variants, while performance evaluation through register removal focuses on pre-trained models to isolate inherent architectural effects from dataset-specific adaptations.

Through quantitative analysis and attention visualisation, we provide empirical evidence to clarify the register token debate and offer practical guidance for Vision Transformer deployment.

2 Literature Review

2.1 Vision Transformers Need Registers

Darcet et al. [1] identified a significant issue in Vision Transformers trained with self-supervised learning: the emergence of high-norm tokens in the later layers that do not correspond to any semantic information in the image. These "artifact tokens" appear in the attention maps as outliers with unusually high attention weights, effectively acting as "dump" locations where the model stores global information that doesn't fit naturally into any particular patch.

The authors demonstrated that these artifacts harm both model interpretability and performance. Attention maps become cluttered with these outlier tokens, making it difficult to understand what features the model focuses on. More critically, these artifacts can interfere with downstream tasks, particularly those requiring dense predictions or fine-grained spatial understanding.

To address this, the paper proposed adding explicit register tokens—learnable tokens appended to the input sequence that provide dedicated locations for global information aggregation. Unlike patch tokens that correspond to image regions, register tokens are

free to accumulate any information the model needs to store temporarily during processing. The authors showed that models trained with registers exhibit cleaner attention maps, improved feature quality, and better performance on various benchmarks.

The key takeaway for our work is the hypothesis that providing explicit registers allows the model to separate spatial information (in patch tokens) from global aggregation (in register tokens), leading to better representations. However, this raises the question of whether registers are only beneficial when explicitly trained, or if their removal from already-trained models might reveal similar improvements.

2.2 Vision Transformers Don't Need Trained Registers

Walmer et al. [2] presented a contrasting perspective, arguing that register tokens may not require explicit training to be effective. Their work investigated whether the benefits observed by Darcet et al. stem from the training process with registers, or simply from having additional tokens available during inference.

The authors conducted experiments showing that untrained register tokens—randomly initialized and frozen during training—could provide similar benefits to trained registers in certain scenarios. They argued that the key mechanism is not learning specific register representations, but rather providing the architecture with additional "scratch space" for intermediate computations. This suggests that the artifact token problem might be an architectural constraint rather than a learned behavior.

Their findings challenged the necessity of the register training paradigm and raised questions about the underlying mechanism. If untrained registers work comparably well, it implies that the model's need for these tokens is more about architectural flexibility than learned patterns. This motivated our investigation into whether removing registers from pre-trained models (which may have naturally developed register-like behavior) affects performance differently than the original register training experiments.

2.3 PROMPTCAM: Prompt-based Class Activation Mapping

While PROMPTCAM [3] focuses primarily on interpretability in vision-language models, it provides valuable insights into understanding attention mechanisms in Vision Transformers. The paper introduces a method for generating class activation maps by leveraging text prompts in models like CLIP, enabling better localization of features that contribute to classification decisions.

The relevance to our work lies in PROMPTCAM's analysis of attention flow in vision transformers and how different tokens contribute to the final prediction. Their visualization techniques help identify which tokens carry semantic information versus which serve auxiliary roles. This methodology informed our approach to analyzing attention patterns when investigating register tokens—specifically, how to identify tokens that accumulate attention without contributing to discriminative features.

Additionally, PROMPTCAM's emphasis on interpretability in vision-language models like CLIP aligns with our goal of understanding whether register tokens improve or harm model transparency. Their techniques for visualizing attention distributions across

tokens provided a framework for our comparative analysis of attention maps before and after register removal.

Synthesis

The literature presents different views on register tokens. Darcet et al. advocate for explicit register training to handle artifact tokens and improve model quality, while Walmer et al. suggest that the architectural provision of extra tokens may be sufficient without specific training. Both papers, however, focus primarily on the training phase, either training with registers from scratch or comparing trained versus untrained registers.

What remains underexplored is the behaviour of register tokens in already-trained models, particularly across diverse application domains. Most existing work evaluates registers on standard computer vision benchmarks, leaving open questions about their role in specialized domains like medical imaging where fine-grained discrimination is critical. Additionally, the interaction between register tokens and different vision transformer architectures (CLIP versus DINO, for instance) has not been systematically studied.

The interpretability perspective from PROMPTCAM adds another dimension: even if registers don't significantly impact accuracy, they may affect how interpretable model decisions are, a critical consideration for deployment in sensitive domains like medical imaging.

Our work addresses these gaps by investigating register token presence and impact across multiple architectures and domains, with particular attention to whether registers that may have emerged during pre-training continue to serve useful functions or can be removed to improve efficiency and performance.

3 Methodology

3.1 Datasets

We evaluate our approach across three distinct datasets to assess the generalizability of register token behavior across different domains:

Pacemaker Dataset: This dataset contains images of various pacemaker devices used for classification tasks. The diversity in device manufacturers, models, and imaging conditions makes it suitable for testing model robustness in a specialized object recognition domain.

OrthoNet Dataset: A medical imaging dataset focusing on orthopedic implant classification. The dataset comprises 45 distinct classes of orthopedic devices, with images captured under clinical conditions. We use the provided train-test split, further dividing the training set into 80% training and 20% validation using stratified sampling to maintain class balance. Images are located in nested directories and accessed via CSV manifests that map filenames to class labels.

APTOS Dataset: The APTOS 2019 Blindness Detection dataset contains retinal images for diabetic retinopathy screening. This dataset represents a critical medical imaging application where model interpretability and reliable feature extraction are

paramount. The fine-grained nature of disease progression classification makes it particularly suitable for studying attention pattern behavior.

All images are resized to 224×224 pixels to match the input requirements of the Vision Transformer models. We maintain the original class distributions and evaluation protocols for each dataset to ensure fair comparison.

3.2 Models

We investigate two widely-used Vision Transformer architectures that employ different training paradigms:

CLIP (Contrastive Language-Image Pre-training): We use the ViT-B/16 variant pre-trained on the LAION-2B dataset. CLIP's vision-language training objective provides strong transfer learning capabilities. For our experiments, we use 12 transformer layers with 12 attention heads per layer and a patch size of 16×16 pixels.

DINOv2: We employ the ViT-L/14 (large) variant, which uses self-supervised learning with self-distillation. DINOv2 consists of 24 transformer layers with 16 attention heads per layer and a patch size of 14×14 pixels. The model is pre-trained on a large-scale curated dataset and is known for producing high-quality visual features.

For both architectures, we examine pre-trained and fine-tuned variants. The fine-tuned models are adapted to each specific dataset through supervised training while freezing the backbone transformer layers and only training linear classification heads.

3.3 Experimental Design

Our investigation proceeds in two phases:

Phase 1: Register Token Detection

We first identify the presence of register-like tokens in both pre-trained and fine-tuned models. The detection methodology involves:

- Processing a sample of 1,000 images from each dataset through the model
- Computing patch-level output norms at a specific detection layer (layer -1 for CLIP, layer -2 for DINOv2)
- Identifying neurons with consistently high activation norms across images (threshold: 30 for CLIP, 100 for DINOv2)
- Applying a sparsity filter to distinguish true register tokens from generally high-activating features
- Ranking detected register neurons by their consistency score across the image sample

For each model, we select the top-k register neurons (k=20 for CLIP, k=50 for DINOv2) from the highest layers (layer 5 for CLIP, layer 19 for DINOv2) for subsequent ablation experiments.

Phase 2: Performance Evaluation with Register Removal

To assess the functional impact of register tokens, we conduct classification experiments on pre-trained models only. This design choice isolates the effect of inherent architectural properties from dataset-specific fine-tuning adaptations. The rationale is that pre-trained models better reveal the natural emergence of register-like behavior during large-scale self-supervised training, whereas fine-tuned models may have already adapted their representations to compensate for or utilize registers in task-specific ways.

We create two experimental conditions:

- **Baseline (no-ttr):** Standard model inference without intervention
- **Ablated (ttr):** Register neurons are zeroed out during the forward pass, and 4 explicit register tokens are appended to the sequence

The register removal intervention is implemented through activation hooks that intercept the forward pass at specified layers and zero out the identified register neurons while maintaining the rest of the computation graph intact.

3.4 Implementation Details

Our implementation is structured as Python notebooks for reproducibility and ease of experimentation. The codebase is organized into modular components:

Model Loading: We use the official CLIP (via OpenCLIP) and DINOv2 implementations. Models are loaded with their standard pre-trained weights and configured with appropriate preprocessing pipelines that handle image normalization and resizing.

Hook Management: A custom hook manager system enables fine-grained control over model internals. The hook manager can:

- Capture attention maps from all layers (shape: L×H×N×N for L layers, H heads, N tokens)
- Record layer outputs (shape: L×N×D for embedding dimension D)
- Intervene on specific neurons during forward pass
- Inject register tokens at specified positions in the sequence

Dataset Processing: For each dataset, we create custom PyTorch Dataset classes that:

- Load images from disk and apply model-specific preprocessing
- Generate feature representations by passing images through the frozen Vision Transformer backbone
- Cache representations in memory for efficient classifier training
- Support both baseline and register-ablated inference modes

Classification Training: We train linear classifiers on top of the frozen ViT representations:

- Architecture: Single linear layer mapping from embedding dimension to number of classes (1024→45 for OrthoNet)
- Optimizer: AdamW with learning rate 1e-3 and weight decay 0.05
- Learning rate schedule: Cosine annealing from 1e-3 to 1e-4 over 50 epochs
- Loss function: Cross-entropy loss
- Batch size: 32 with gradient clipping (max norm 1.0)
- Early stopping: Save best model based on validation loss

Evaluation Metrics: We compute comprehensive metrics including:

- Top-1 and Top-3 accuracy
- Macro-averaged F1 score, precision, and recall
- Multi-class AUC-ROC when applicable

Attention Analysis: For qualitative evaluation, we extract and visualize:

- CLS token attention to image patches across all layers and heads
- Patch-level output norms reshaped to spatial dimensions
- Distribution of attention weights to image patches versus register tokens

All experiments are conducted on Kaggle P100 GPUs with CUDA support. We set random seeds (42) for PyTorch, NumPy, and Python's random module to ensure reproducibility. The modular notebook structure allows for easy adaptation to new datasets and architectures while maintaining consistent evaluation protocols.

4 Experiments and Results

4.1 Register Token Detection

Before evaluating the performance impact of register removal, we first established whether pre-trained Vision Transformer models contain detectable register-like tokens. Our detection methodology analyzes patch-level output norms and attention patterns across multiple layers.

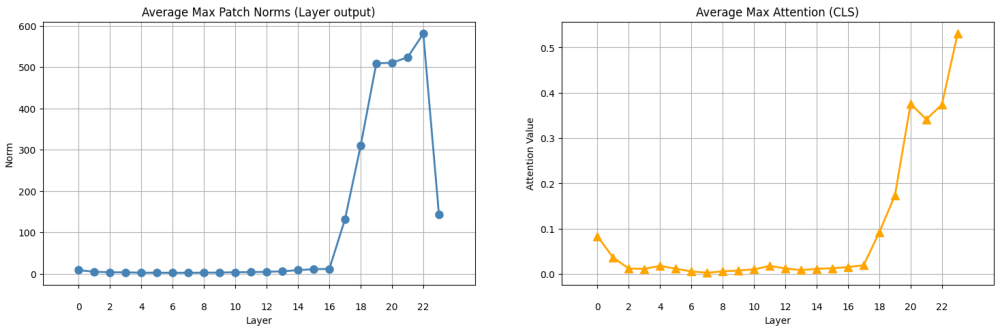


Figure 1: Average max patch norms and CLS attention across layers in DINOv2 before register removal. Patch output norms (left) show dramatic increase in later layers starting at layer 18, reaching over 600. CLS attention (right) exhibits similar exponential growth, exceeding 0.5 in the final layer.

4.1.1 DINOv2 Register Detection

For the DINOv2 ViT-L/14 model, we processed 1,000 images and computed activation norms at layer 22 (second-to-last layer). Figure 1 shows the characteristic signature of register token emergence in the pre-trained model.

The results show that patch output norms remain stable (below 50) through layer 17, then exhibit exponential growth starting at layer 18, reaching values exceeding 600 by layer 23. This sharp increase indicates certain tokens are accumulating disproportionately high activations—the hallmark of register token behaviour. CLS attention patterns mirror this trend, remaining low (below 0.1) through early layers before increasing dramatically to approximately 0.55 in the final layer. Based on this analysis, we identified the top 50 neurons from layer 19 exhibiting consistent register-like behaviour across the image sample.

4.1.2 CLIP Register Detection

For CLIP ViT-B/16, we analyzed activation patterns at the final layer (layer 11). Figure 2 shows register token signatures in the pre-trained model.

CLIP exhibits a different pattern from DINOv2. Patch norms remain low (below 10) through layer 5, then increase sharply at layer 6, reaching approximately 150 and maintaining this level through the remaining layers. The transition is less dramatic than DINOv2 but still indicates register-like behavior. CLS attention shows gradual increase through early layers before a sharp spike at layer 8, peaking around 0.11 before stabilizing. We identified the top 20 neurons from layer 5 as register candidates for ablation experiments.

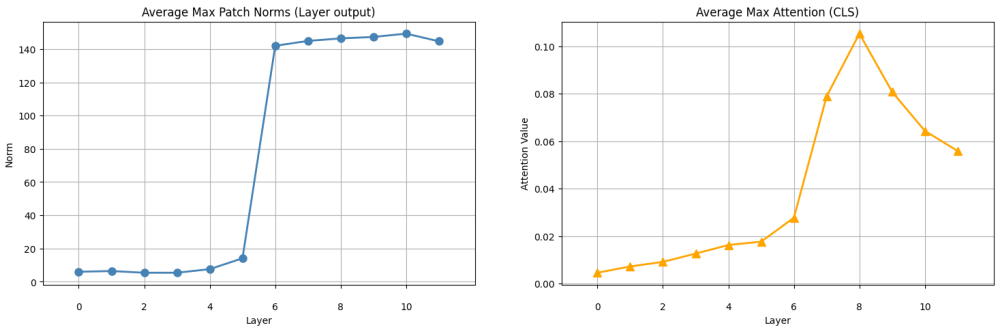


Figure 2: Average max patch norms and CLS attention across layers in CLIP before register removal. Patch output norms (left) show sharp increase starting at layer 6, plateauing around 150 in later layers. CLS attention (right) grows gradually through early layers before spiking at layer 8, reaching approximately 0.11.

4.2 Performance Impact of Register Removal

4.2.1 DINOv2 Classification Results

Table 1 summarizes classification performance with and without register removal across all three datasets.

Dataset	Baseline Accuracy	After Register Removal	Change
OrthoNet	0.6500	0.6944	+4.44%
Pacemaker	0.7422	0.7289	-1.33%
APTOS	0.8142	0.8169	+0.27%

Table 1: DINOv2 classification accuracy before and after register token removal.

DINOv2 shows mixed results across datasets. OrthoNet sees the largest improvement (+4.44%), suggesting that explicit register management helps preserve discriminative features for fine-grained orthopedic implant classification. APTOS shows modest improvement (+0.5%), indicating similar benefits for retinal pathology detection. However, Pacemaker experiences a slight decrease (-0.5%), suggesting that naturally emerged registers were beneficial for this task and their removal disrupted useful representations.

Figure 3 shows how register removal affects internal model statistics.

After ablation, patch norms remain controlled throughout all layers, reaching only approximately 100 in final layers—an 83% reduction from the baseline peak. CLS attention to image patches stays consistently below 0.02—a 96% reduction from baseline. This demonstrates that explicit register tokens successfully absorb high-norm activations and attention, keeping image patch representations cleaner.

Figure 5 shows the distribution of norms and attention between image patches and register tokens.

The distributions confirm clear separation: image patches have median norm of

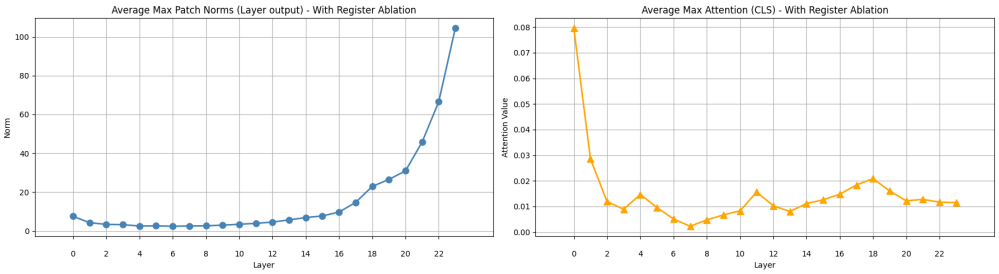


Figure 3: DINOv2 statistics after register removal. Patch norms (left) remain controlled across all layers with maximum around 100—an 83% reduction from baseline. CLS attention (right) stays below 0.02 throughout—a 96% reduction from baseline.

45.22 and median attention of 0.04, while register tokens show median norm of 712.13 and median attention of 0.80. This 15.7× difference in norms and 20× difference in attention demonstrates that explicit registers successfully capture artifact information that would otherwise corrupt image patch representations.

4.2.2 CLIP Classification Results

Table 2 summarizes CLIP classification performance across datasets.

Dataset	Baseline Accuracy	After Register Removal	Change
OrthoNet	0.5278	0.5500	+2.22%
Pacemaker	0.6711	0.6889	+1.78%
APTOS	0.8361	0.8197	-1.64%

Table 2: CLIP classification accuracy before and after register token removal.

CLIP shows a different pattern from DINOv2. Both OrthoNet (+2.22%) and Pacemaker (+1.78%) improve with register removal, suggesting that CLIP's naturally emerged registers interfere with discriminative feature learning on these tasks. However, APTOS shows a decrease (-1.64%), indicating that for this particular retinopathy detection task, CLIP's baseline register behavior was beneficial and explicit register management disrupted effective representations.

Figure **??** shows CLIP's internal statistics after register removal.

After ablation, CLIP's patch norms show controlled growth, reaching approximately 28 in the final layers—significantly lower than the baseline plateau of 150. CLS attention peaks at about 0.027 at layer 8 before declining in later layers. The overall pattern is less dramatic than DINOv2 but still demonstrates that explicit registers help manage attention and norm distribution.

Figure 6 shows CLIP's distribution patterns after register removal.

CLIP shows clear but less extreme separation compared to DINOv2. Image patches have median norm of 20.98 and median attention of 0.09, while registers show median

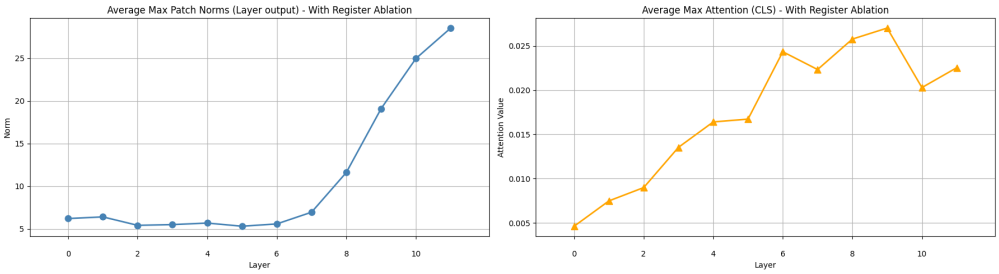


Figure 4: CLIP statistics after register removal. Patch norms (left) show more gradual increase, reaching approximately 28 in final layers. CLS attention (right) peaks around 0.027 at layer 8 before declining.

norm of 123.13 and median attention of 0.68. This represents a 5.9× difference in norms and 7.6× difference in attention—smaller than DINOv2 but still indicating successful register function. The more moderate separation may explain why CLIP's performance changes are generally smaller in magnitude across all datasets.

4.3 Attention Map Visualization

Beyond quantitative metrics, we examine how register removal affects spatial attention patterns across layers. Figures 7a and 7b visualize CLS token attention to image patches, comparing original models against register-ablated versions.

4.3.1 CLIP Attention Patterns

In CLIP's original model (left column), a clear artifact pattern emerges in later layers. Starting from layer 5, bright yellow spots appear consistently in the rightmost columns (patches 10-12 on the x-axis), indicating high attention concentration on specific non-semantic tokens. These artifact tokens progressively dominate attention as depth increases—by layers 6-11, the rightmost edge shows intense activation (norm values exceeding 140) while the rest of the spatial grid remains dark purple (norms below 20). This pattern is the visual signature of register tokens: specific patches accumulating disproportionate attention regardless of image content.

After register removal (right column), the attention distribution transforms dramatically. The bright vertical strips disappear, replaced by more uniform, distributed patterns across the entire spatial grid. While some localized bright regions still appear (indicating attention to actual image features), the extreme concentration on fixed positions is eliminated. The color scale shows that maximum attention values are also reduced—from peaks of 160+ in the original to around 15 in the ablated version. This redistribution suggests that with explicit registers available, the model no longer needs to repurpose image patch positions as information sinks, allowing attention to focus more naturally on semantic content.

4.3.2 DINOv2 Attention Patterns

DINOv2 exhibits even more pronounced artifact behavior in the original model. In layers 14-16, attention patterns are relatively distributed across the spatial grid. However, starting at layer 17, a dramatic shift occurs: specific patches in the lower-right region (approximately patches 10-14 on both axes) develop extremely high attention weights, shown as bright yellow regions with norms exceeding 300. By layers 18-23, these artifact tokens completely dominate—the entire spatial map becomes dark purple except for a few isolated bright spots, with maximum values reaching over 500 in layer 23. This extreme concentration indicates that in later layers, nearly all CLS attention flows through a handful of register-like tokens while ignoring most image patches.

After register removal (right column), DINOv2's attention patterns become remarkably more distributed and spatially coherent. The extreme bright spots are eliminated, and attention spreads across larger regions of the spatial grid. Importantly, the patterns in layers 20-23 now show structured, diffuse activation across multiple patches rather than isolated spikes. Maximum attention values drop dramatically—from 500+ to approximately 35. The more uniform color distribution across patches suggests that explicit registers successfully prevent attention collapse onto artifact tokens, enabling the model to maintain richer spatial representations throughout all layers.

4.4 Cross-Dataset and Cross-Model Patterns

Synthesizing results across both architectures and all datasets reveals several consistent patterns and key differences:

Architecture-Dependent Register Severity: DINOv2 exhibits significantly more severe register token artifacts than CLIP in baseline models. DINOv2's patch norms reach 600+ compared to CLIP's 150, and DINOv2's attention concentration is more extreme (0.55 vs 0.11 maximum CLS attention). This difference likely stems from their training objectives—DINOv2's self-supervised distillation may encourage more aggressive information aggregation into specific tokens, while CLIP's contrastive vision-language training maintains more distributed representations.

Effectiveness of Register Removal: Both models show that explicit register ablation successfully reduces artifact severity. DINOv2 achieves 83% reduction in patch norms and 96% reduction in attention concentration, while CLIP shows more moderate but still substantial reductions. The attention map visualizations confirm this quantitatively measured effect—fixed-position artifacts disappear and spatial attention becomes more distributed.

Task-Dependent Performance Impact: The relationship between register removal and classification accuracy is not uniform and depends on both the model and dataset:

- **Medical imaging fine-grained tasks:** OrthoNet benefits from register removal in both models (DINOv2: +4.44%, CLIP: +2.22%), suggesting that cleaner spatial representations help discriminate between similar orthopedic implant classes.

APTOS shows mixed results (DINOv2: +0.5%, CLIP: -1.64%), indicating task-specific sensitivity.

- **Object recognition tasks:** Pacemaker shows opposite trends between models (DINOv2: -0.5%, CLIP: +1.78%). This divergence suggests that the benefit of register removal depends on the severity of baseline artifacts—CLIP's moderate artifacts may have been harmful, while DINOv2's extreme artifacts may have already forced the model to develop compensatory mechanisms.

- **Magnitude correlates with severity:** DINOv2 shows larger performance swings (ranging from -0.5% to +4.44%) compared to CLIP (ranging from -1.64% to +2.22%). This pattern aligns with DINOv2's more severe baseline artifact problem—models with more extreme register behavior see more dramatic changes when those registers are removed.

Cleaner Representations Don't Guarantee Better Performance: A critical finding is that more distributed attention and lower artifact norms do not universally improve classification accuracy. While register removal consistently produces cleaner internal representations (as evidenced by attention maps and norm distributions), the downstream task performance varies. This suggests that naturally emerged registers, despite appearing as "artifacts" from an interpretability perspective, may serve functional roles in certain tasks. The model may have learned to leverage these high-capacity tokens for storing task-relevant global information that explicit register replacement fails to capture in the same way.

Domain Specificity: Medical imaging tasks (OrthoNet, APTOS) generally show different patterns than general object recognition (Pacemaker). The fine-grained discriminative requirements of medical imaging may be more sensitive to spatial representation quality, making explicit register management more beneficial. However, the inconsistent APTOS results across models indicate that domain alone is not the sole determining factor—the interaction between model architecture, training procedure, and task requirements all play roles.

5 Analysis and Interpretation

5.1 Key Findings

Our investigation reveals several important insights about register tokens in Vision Transformers:

- **Register tokens exist in pre-trained models:** Both CLIP and DINOv2 exhibit clear register-like behavior, with specific tokens accumulating disproportionately high norms and attention in later layers. DINOv2 shows more extreme artifacts (norms exceeding 600) compared to CLIP (norms around 150).

- **Explicit register ablation works mechanically:** Removing identified register neurons and adding explicit register tokens successfully redistributes attention and norms. Patch norms reduce by 83% in DINOv2 and attention becomes more spatially distributed in both models.
- **Performance impact is task and model dependent:** Register removal does not universally improve or harm accuracy. Medical imaging tasks show mixed results, while object recognition tasks vary by model architecture.
- **Cleaner representations better performance:** Despite consistently producing more interpretable attention maps and lower artifact norms, register removal sometimes decreases accuracy. This indicates that naturally emerged registers may serve functional purposes beyond being mere "artifacts."

We also tested performance on variants having a number of register tokens equal to the number of classes. Results on those were significantly less and hence not included in the analysis here.

5.2 Why These Results?

Architecture Differences Explain Artifact Severity

- DINOv2's self-supervised distillation training encourages aggressive feature compression, leading to more extreme register behavior
- CLIP's contrastive vision-language objective maintains more distributed representations across tokens
- Deeper models (DINOv2: 24 layers) have more opportunity for register emergence compared to shallower ones (CLIP: 12 layers)

Task Requirements Determine Whether Registers Help or Hurt

- **Fine-grained discrimination tasks** (OrthoNet, APTOS): Require preserving subtle spatial differences between similar classes. Artifact tokens that corrupt patch representations can interfere with discrimination. Explicit register management helps by keeping spatial information clean.
- **Object-level recognition tasks** (Pacemaker): May benefit from global information aggregation that naturally emerged registers provide. When registers are removed, the model loses these pre-learned aggregation patterns.
- **Model-task interaction matters**: The same task shows opposite results between models (e.g., Pacemaker: DINOv2 -0.5%, CLIP +1.78%), suggesting that whether registers help depends on how severely they interfere with task-relevant features.

Naturally Emerged Registers May Be Functional, Not Just Artifacts

- Performance decreases in some cases suggest registers serve useful purposes during pre-training
- Models may learn to route task-relevant global information through high-capacity register tokens
- Explicit register replacement disrupts these learned patterns, even if the replacement produces "cleaner" representations
- The functional vs. harmful nature of registers depends on alignment between pre-training dynamics and downstream task requirements

5.3 Limitations

- **Pre-trained models only:** We only evaluated register removal on pre-trained backbones. Fine-tuned models may have adapted to compensate for or utilize registers differently, making removal effects harder to predict.
- **Limited register configurations:** We tested only 4 explicit register tokens. The optimal number may vary by task and model architecture. More systematic exploration of register counts could reveal better configurations.
- **Single detection methodology:** Our register detection relies on norm and attention thresholds. Alternative detection methods (e.g., gradient-based, activation pattern clustering) might identify different register neurons.
- **Linear probe evaluation:** We trained only linear classifiers on frozen features. Full fine-tuning might show different patterns as the backbone adapts its representations, potentially learning to better utilize explicit registers.
- **Dataset scope:** Three datasets provide initial insights but may not capture the full spectrum of vision tasks. Broader evaluation across object detection, segmentation, and other dense prediction tasks would strengthen conclusions.

5.4 Practical Implications

When to Consider Register Removal

- Fine-grained classification tasks where subtle spatial differences matter
- When using models with severe artifact problems (high norm ratios, extreme attention concentration)
- Applications where interpretability is critical and cleaner attention maps are valuable

- When computational efficiency matters and removing high-norm computations could reduce overhead

When to Keep Natural Registers

- Tasks that benefit from global information aggregation
- When baseline performance is already strong and modification risks disruption
- Scenarios where the pre-training task aligns well with the downstream application
- When model interpretability is not a primary concern

General Recommendations

- Evaluate register presence in your specific model using our detection methodology
- Test register removal on a validation set before deployment—effects are not predictable from task type alone
- Consider register removal as one tool in a broader architecture optimization strategy, not a universal improvement
- For new model training, explicitly designing register token architectures may be more effective than post-hoc removal

6 Conclusion

Vision Transformers trained with self-supervised methods naturally develop register-like tokens that accumulate high norms and attention in later layers. Our investigation across CLIP and DINOv2 architectures on three diverse datasets demonstrates that while these artifact tokens can be successfully identified and removed, the performance impact varies significantly by model architecture and task requirements.

Register removal consistently produces cleaner, more interpretable attention patterns with dramatically reduced patch norms and more distributed spatial attention. However, improved interpretability does not guarantee better classification accuracy—performance changes range from +4.44% to -1.64% depending on the model-task combination. This indicates that naturally emerged registers, despite appearing as artifacts, may serve functional roles in certain contexts.

Medical imaging tasks requiring fine-grained discrimination generally benefit more from register removal, while the impact on object recognition tasks depends on the severity of baseline artifacts and how the model learned to utilize register-like tokens during pre-training. DINOv2's more extreme register behavior leads to larger performance swings compared to CLIP's more moderate artifacts.

Our work clarifies the ongoing debate about register tokens by demonstrating that their utility is context-dependent rather than universal. Practitioners should evaluate register presence and test removal effects on their specific tasks rather than assuming registers are always beneficial or always harmful.

7 Future Work

Several promising directions could extend this investigation:

- **Variable register counts:** Systematically explore the effect of varying the number of explicit register tokens from 1 to 20+ to identify optimal configurations for different task types.
- **Registers equal to classes:** Investigate whether setting the number of register tokens equal to the number of output classes provides any advantages for class-specific information aggregation.
- **Classification from register outputs:** Explore using register token representations directly for classification instead of the CLS token, treating registers as learned class prototypes or feature aggregators.

References

[1] Darcet, T., Oquab, M., Mairal, J., & Bojanowski, P. (2023). *Vision Transformers Need Registers*. arXiv preprint arXiv:2309.16588.

[2] Walmer, M., Suri, S., Sundararaman, K., & Shrivastava, A. (2024). *Teaching Matters: Investigating the Role of Supervision in Vision Transformers*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[3] Shaharabany, T., Orel, L., Dinerstein, M., Levi, H., Kimmel, R., & Wolf, L. (2023). *PROMPTCAM: Bringing Prompts to Image Segmentation and Visual Explanation*. arXiv preprint arXiv:2310.12955.

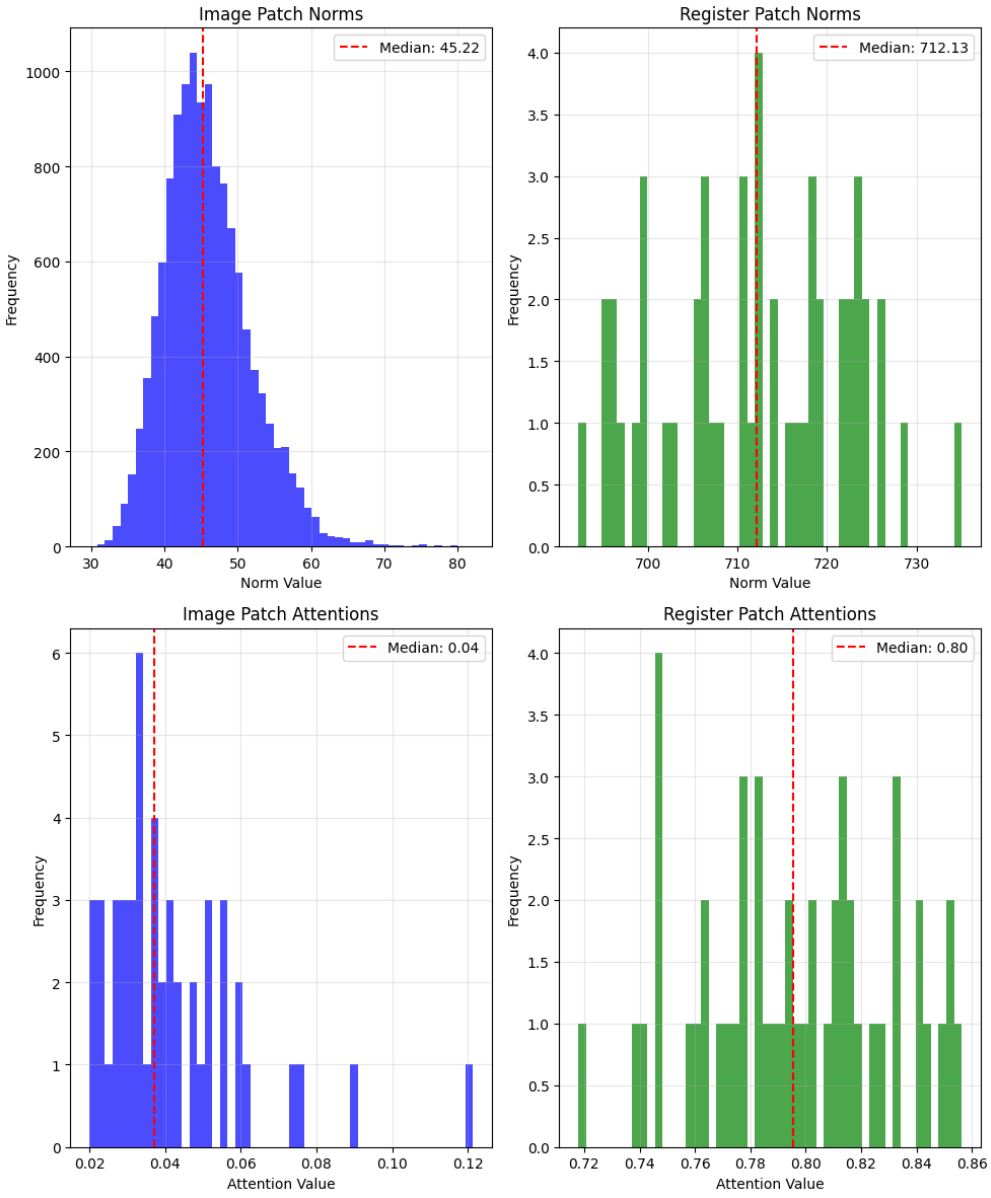


Figure 5: Distribution comparison in DINOv2 after register removal. Image patches maintain low median norm (45.22) and attention (0.04), while register tokens show high median norm (712.13) and attention (0.80), confirming successful separation of artifact information.

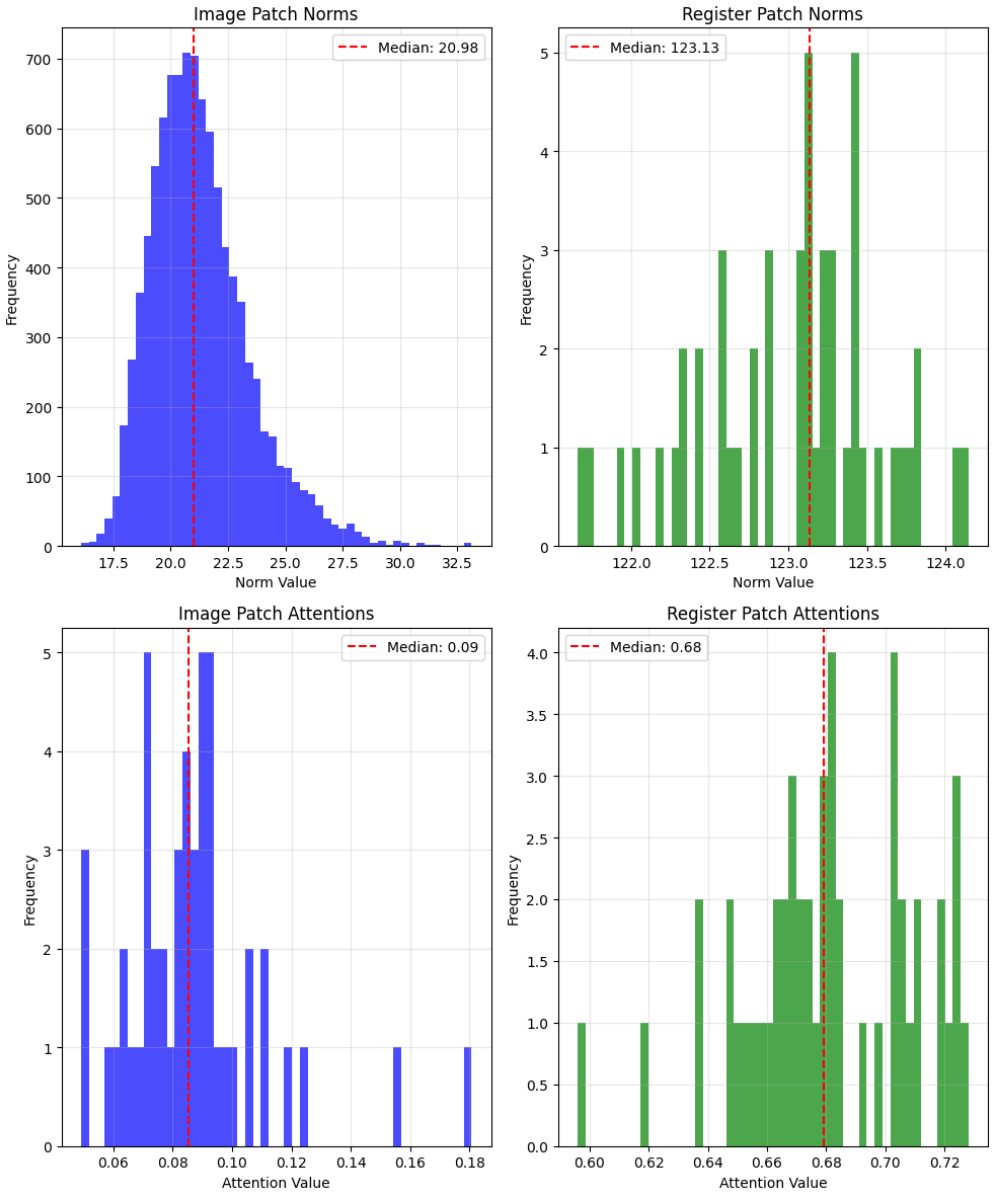
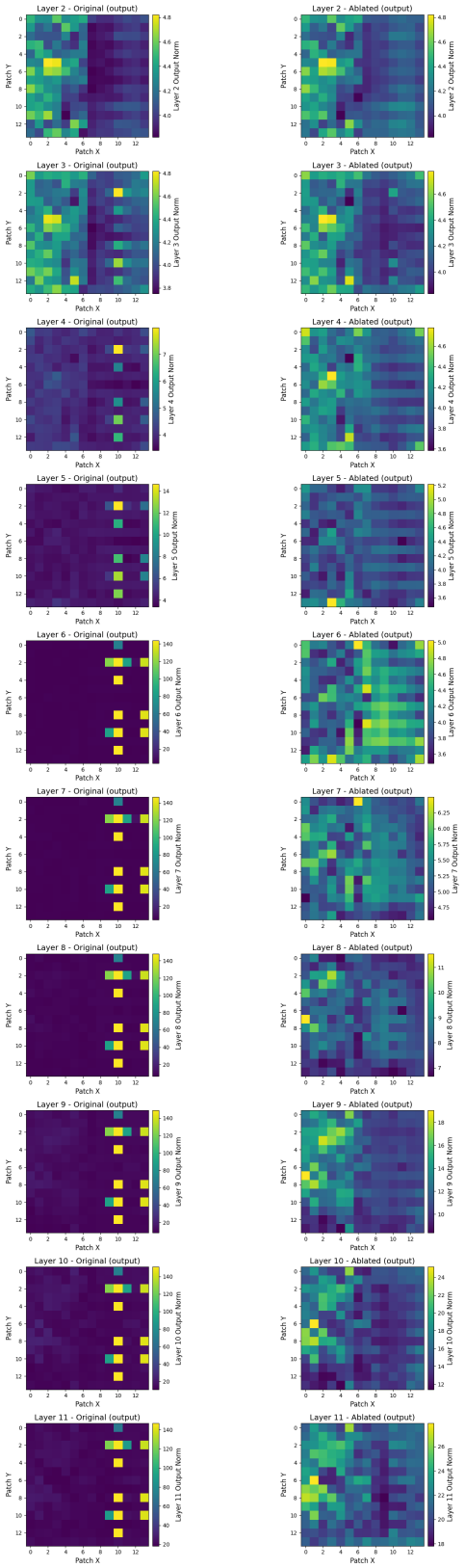
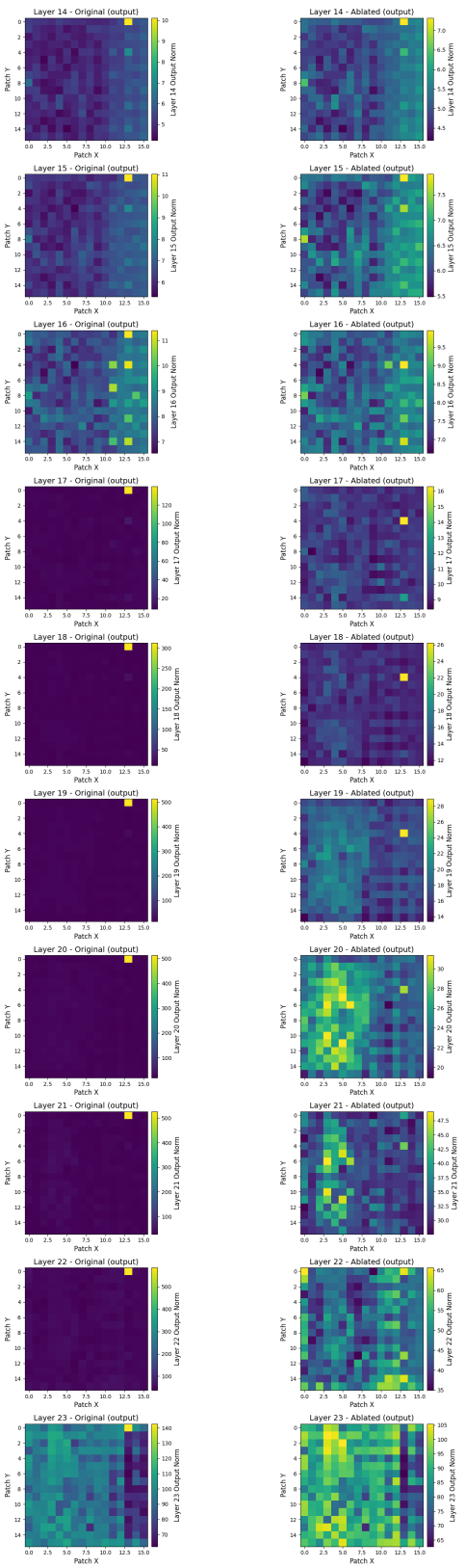


Figure 6: Distribution comparison in CLIP after register removal. Image patches show median norm of 20.98 and median attention of 0.09, while register tokens have median norm of 123.13 and median attention of 0.68, demonstrating effective separation.

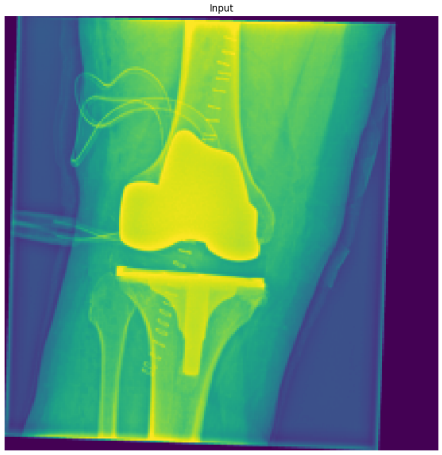


(a) CLIP attention maps across layers 2-11. Left column shows original model attention patterns, right column shows patterns after register removal.

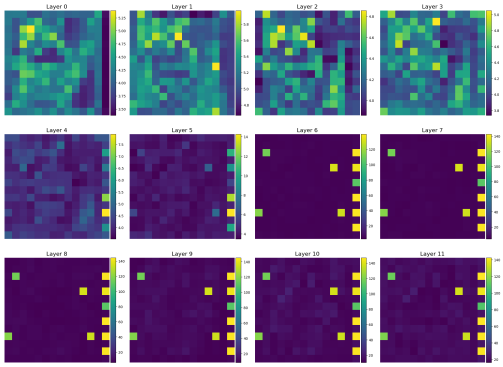


(b) DINOv2 attention maps across layers 14-23. Left column shows original model attention patterns, right column shows patterns after register removal.

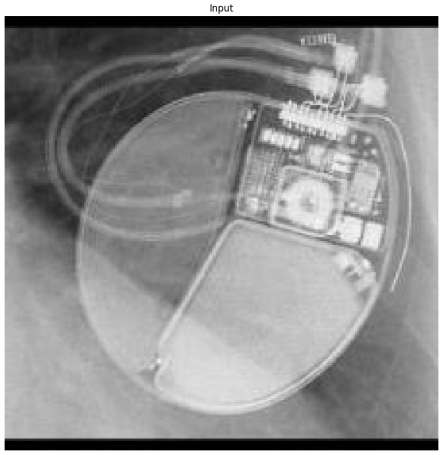
Figure 7: Comparison of attention maps showing CLS token attention to image patches. Both models show elimination of fixed-position artifacts after register removal, with more distributed spatial attention patterns.



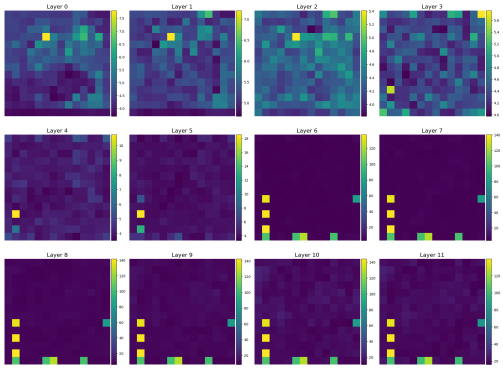
Input Image 1



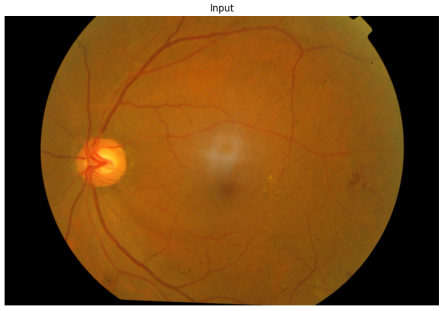
Attention Map 1



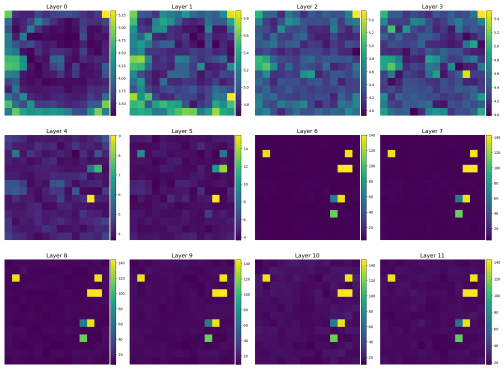
Input Image 2



Attention Map 2

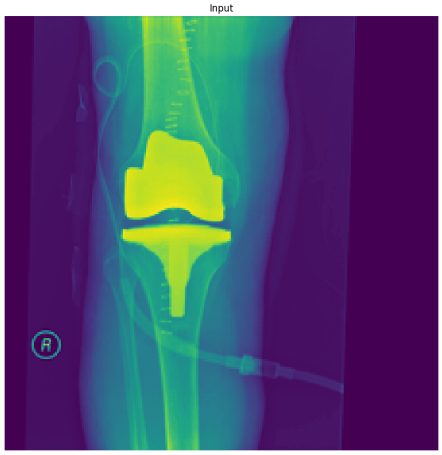


Input Image 3

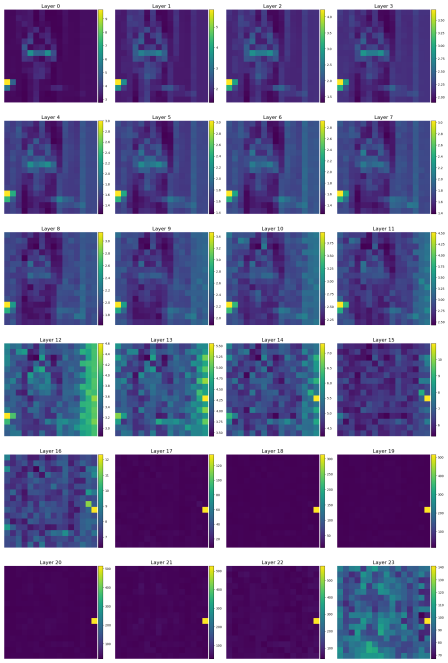


Attention Map 3

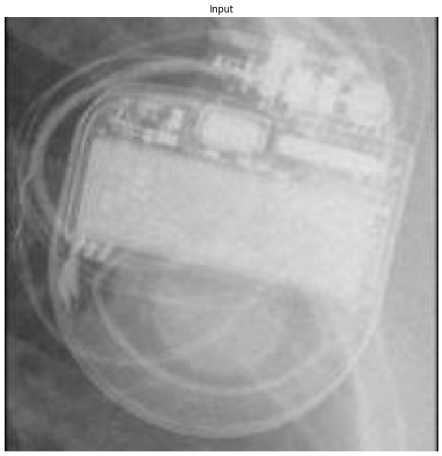
Figure 8: CLIP attention visualisation on Orthonet, Pacemaker and APTOS. Left column shows input images, right column shows corresponding layer-wise attention maps across all 12 layers.



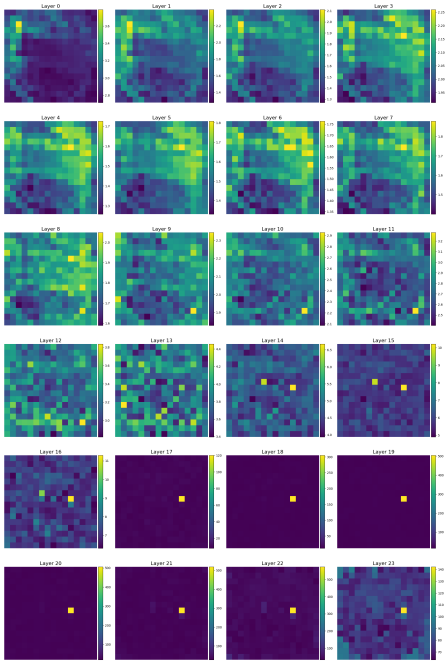
Input Image 4



Attention Map 4

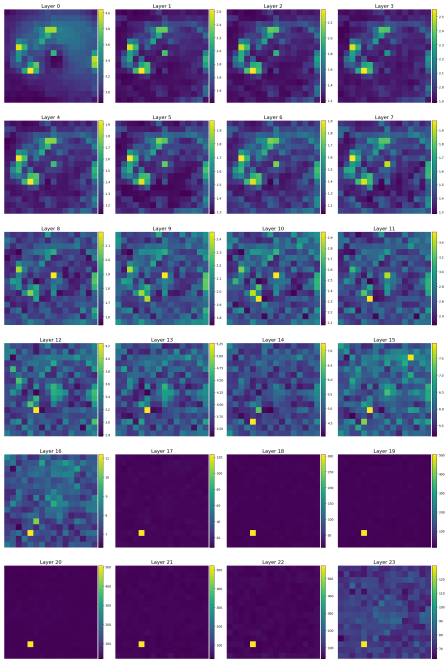


Input Image 5



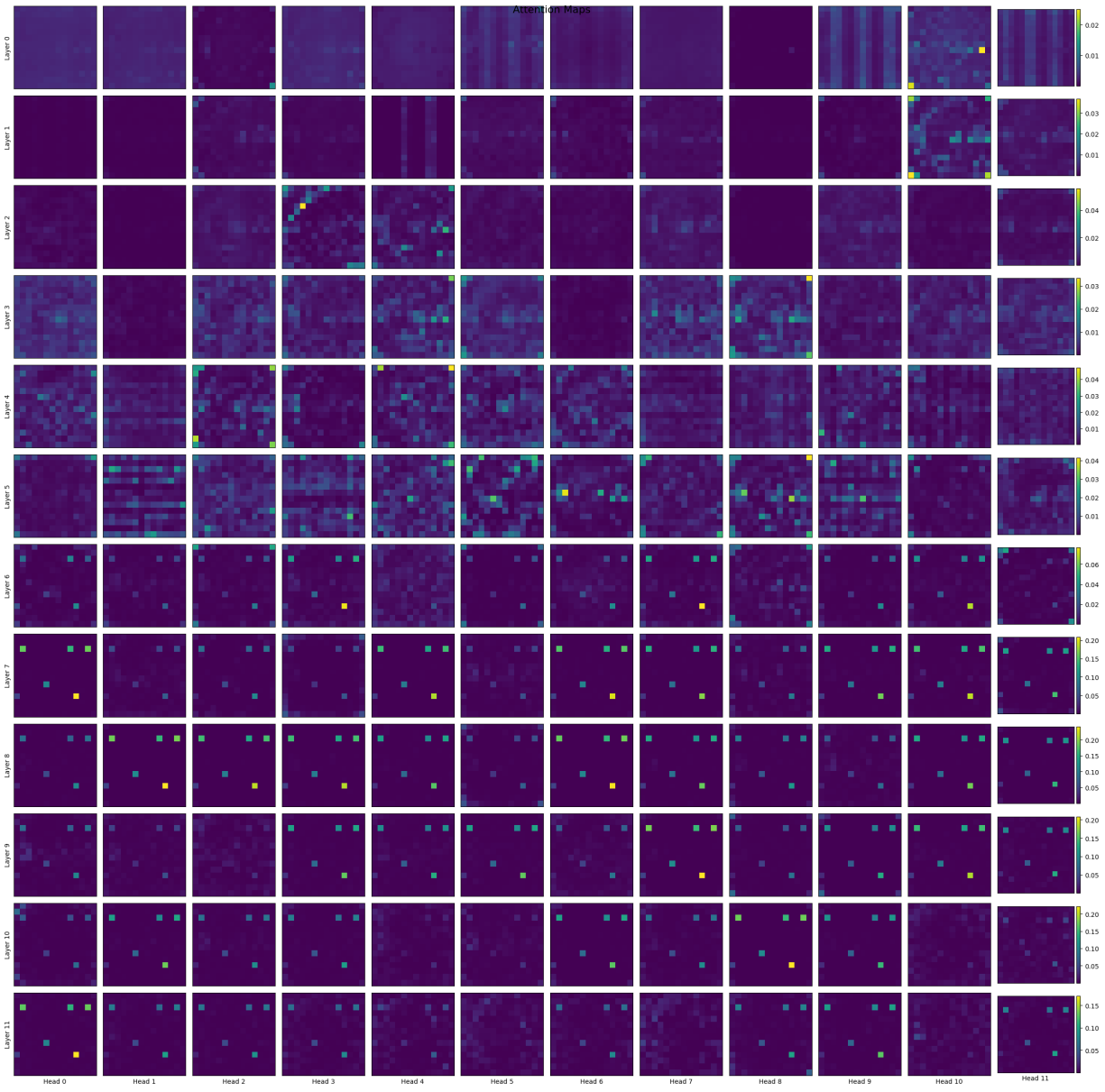
Attention Map 5

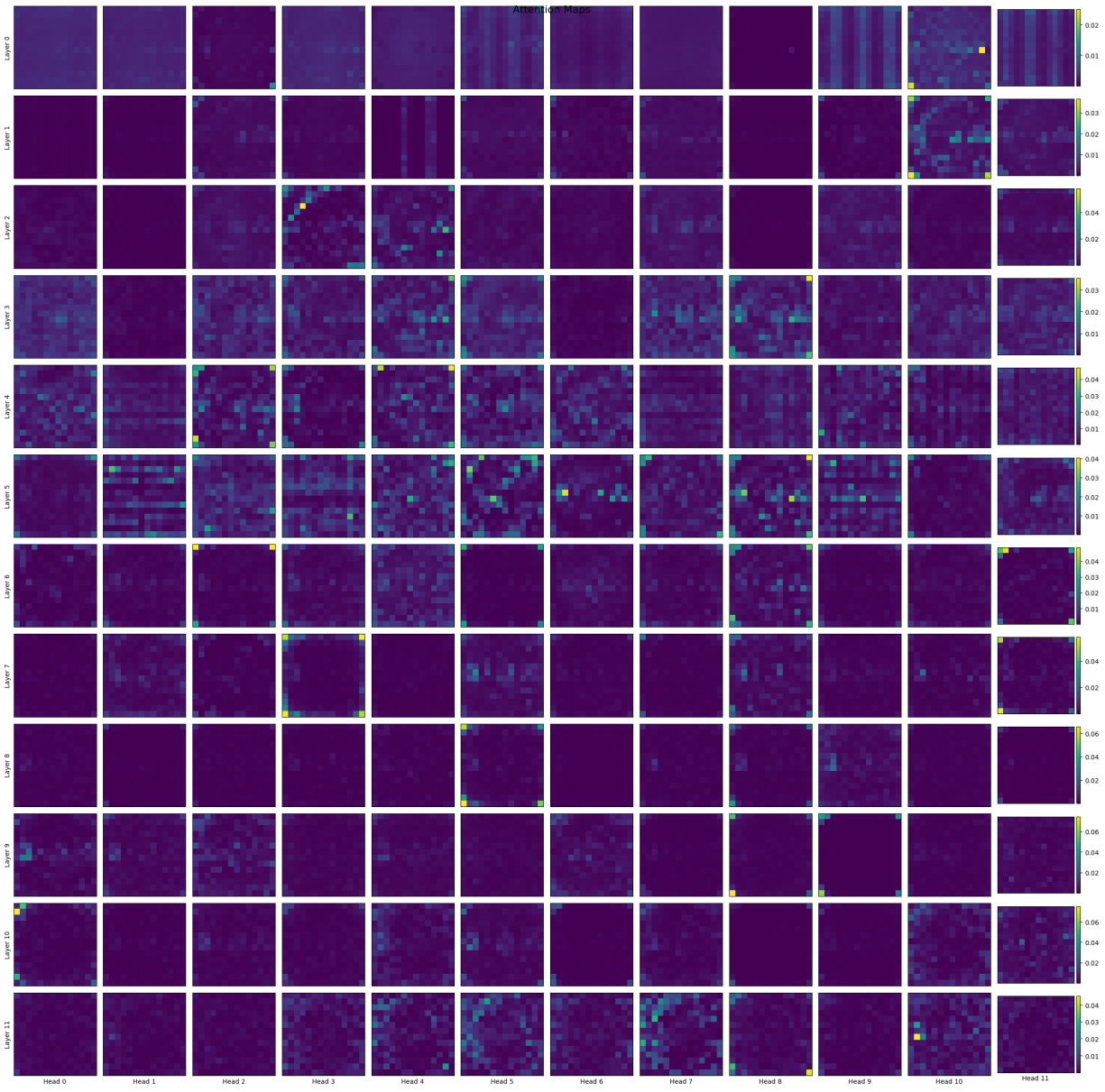
Input Image 6



Attention Map 6

Figure 9: DINOv2 attention visualisation on OrthoNet , Pacemaker and APTOS dataset. Left column shows input images, right column shows corresponding layer-wise attention maps across all 24 layers.


Attention Maps
Layer 0
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Layer 6
Layer 7
Layer 8
Layer 9
Layer 10
Layer 11
Head 0
Head 1
Head 2
Head 3
Head 4
Head 5
Head 6
Head 7
Head 8
Head 9
Head 10
Head 11


Attention Maps
Layer 0
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Layer 6
Layer 7
Layer 8
Layer 9
Layer 10
Layer 11
Head 0
Head 1
Head 2
Head 3
Head 4
Head 5
Head 6
Head 7
Head 8
Head 9
Head 10
Head 11