DO SELF-SUPERVISED VISION MODELS LEARN WHAT EXPERTS SEE?

Desmond Choy, Leong Kay Mei

NUS-ISS, National University of Singapore, Singapore 119615

ABSTRACT

Self-supervised vision models achieve strong downstream performance, but high classification accuracy alone does not reveal whether those models attend to the same visual evidence that human experts consider diagnostically important. This project studies that question in the WikiChurches setting, where expert bounding boxes identify architectural features such as arches, windows, towers, and facade elements that matter for style recognition.

We evaluate seven vision models across self-distillation, masked autoencoding, multimodal contrastive pretraining, and a supervised CNN baseline, then measure attention alignment against 631 expert boxes on 139 annotated church images using IoU, Coverage, MSE, KL divergence, and EMD.

The study is organised around three linked questions: (1) How well frozen models align with expert-marked regions; (2) How Linear Probe, LoRA, and Full fine-tuning change that alignment; (3) Whether individual attention heads exhibit descriptive specialisation for different architectural features.

For Q1, frozen expert-aligned attention is present but highly model-family dependent: DINOv3 leads the default-method benchmark on IoU@90, Coverage, KL, and EMD, and uniquely clears all calibrated continuous baselines across MSE, KL, and EMD. Base SigLIP is a competitive overlap result at its best mean-attention layer (IoU@90 $= 0.074$ at `layer8`, ahead of MAE and CLIP), but SigLIP2 remains weaker on overlap despite the lowest MSE; the two should not be collapsed into one family-level frozen result. For Q2, fine-tuning moves attention unevenly across families: CLIP gains the most (IoU $0.0181 \rightarrow 0.0745$, Cohen's $d \approx 1.0$) with gains concentrated on Gothic and Romanesque features; MAE's largest single-style gain is on Renaissance, driven by pediment geometry; the SigLIP variants improve from weaker `layer11` baselines; and the DINO family preserves its already-strong frozen alignment. Models with different pretraining objectives converge on the same structurally easy images rather than complementary subsets, with DINOv3 frozen IoU predicting per-image CLIP Δ at Pearson $r = +0.677$. For Q3, per-head specialisation is sparse and family-shaped: DINO-family dominant heads stay stable across adaptation (DINOv3 `layer10/head8`, DINOv2 `layer11/head11`), MAE is partly reshaped, and CLIP reorganises from an early frozen head (`layer4/head5`) toward late-layer adapted heads. The clearest head-feature alignments concentrate on larger architectural structures such as portals, arches, and rose windows.

Keywords: self-supervised learning, attention alignment, vision transformers, architectural recognition

1. INTRODUCTION

A vision model can be correct for the wrong visual reasons. In architectural style recognition, a model that classifies a church as Gothic because it attends to pointed arches and flying buttresses is qualitatively different from one that succeeds because it exploits background regularities or photographer bias unrelated to expert reasoning. Accuracy alone cannot distinguish those cases.

WikiChurches [1] provides a strong evaluation setting because it pairs fine-grained architectural-style labels with expert bounding boxes marking characteristic visual features, making it possible to compare model attention directly against human expert targets rather than inferring plausibility from class predictions alone.

This report addresses three linked research questions. **Q1**: Do frozen SSL and baseline vision models attend to the same architectural regions that experts mark as diagnostically important? **Q2**: How does attention change after adaptation (Linear Probe, LoRA, Full fine-tuning), and does strategy matter? **Q3**: Do individual attention heads exhibit descriptive specialisation for different architectural features, and do dominant heads shift across variants?

Taken together, these questions shift the project from a visualisation exercise into a domain-grounded evaluation study of what different pretraining and adaptation choices encourage models to attend to.

2. LITERATURE REVIEW

This section cites papers that carry a concrete design choice in this repository: the dataset, the evaluated model families, the attention-alignment method, the fine-tuning comparison, or the Q3 head-level unit of analysis.

2.1. Attention Alignment and WikiChurches

WikiChurches [1] is the project anchor because it provides both architectural style labels and expert bounding boxes for characteristic building parts. Those boxes turn the project from a generic interpretability demo into a domain-grounded alignment test: do model heatmaps land on the same architectural evidence that experts mark?

Caron et al. [2] motivate the DINO-family part of the study by showing that self-supervised ViT attention can produce object-like masks without localisation supervision. Oquab et al. [3] and Siméoni et al. [4] are direct model sources for DINOv2 and DINOv3, which matter because DINOv3 is the strongest frozen model in the Q1 artifacts and the DINO family is the main stability case in Q2.

Chung et al. [5] are the closest methodological neighbour: they compare ViT attention maps with expert medical annotations rather than generic object masks. This report makes the same domain-specific move, but for architectural heritage and across a wider model and adaptation matrix. Abnar and Zuidema [6] are included because this repo implements attention rollout as an alternative to single-layer CLS attention. Chefer et al. [7] provide the key guardrail: attention visualisations support alignment analysis but should not be treated as full causal explanations.

2.2. Fine-Tuning and Attention Shift

Q2 asks whether adaptation changes where a model looks. Kumar et al. [8] justify the Linear Probe vs. Full fine-tuning contrast by showing that full fine-tuning can distort pretrained features rather than merely improving a classifier head. Biderman et al. [9] motivate the LoRA middle case: parameter-efficient adaptation can learn less and forget less than full fine-tuning. Li et al. [10] argue that attention patterns carry transfer signal; this repo asks the stricter question of whether those changed patterns move toward expert architectural evidence. Two further papers shape the expectation that different pretraining regimes do not respond to adaptation in the same way. Walmer et al. [11] compare supervised, image-text contrastive, masked-autoencoding, and self-distillation training and show that supervision choice produces qualitatively different ViT attention, motivating per-family Q2 reporting rather than pooled deltas. Park et al. [12] compare contrastive learning with masked image modelling and find that the two objectives induce different attention patterns - contrastive features capturing longer-range global structure while masked-image-modelling features remain more local - a distinction that informs the Q2 reading that CLIP-family and MAE improvements concentrate on different image subsets.

2.3. Per-Head Specialisation

Q3 should stay descriptive. Voita et al. [13] are the useful precedent for the idea that a small subset of attention heads carries interpretable, task-relevant behaviour. Li et al. [14] provide the vision-specific counterpart by analysing head importance and attention patterns in ViTs. This report uses those papers to justify a narrower question: whether individual heads in DINOv2, DINOv3, MAE, and CLIP align more strongly with particular expert-marked architectural features, without claiming that a high-ranking head causally explains the model's final prediction.

3. PROPOSED APPROACH

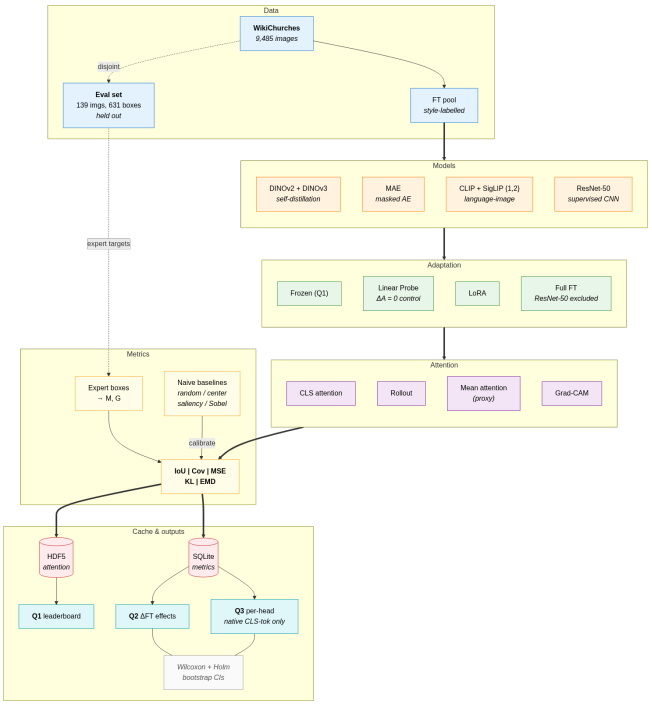


Fig. 1: End-to-end pipeline. The 139-image expert-annotated evaluation set is disjoint from the fine-tuning pool. Linear Probe is a zero-Δ control (backbone frozen); ResNet-50 is excluded from Q2. Attention extraction depends on model family; Q3 is restricted to native CLS-token models. Five alignment metrics are computed against expert boxes (M binary, G Gaussian) and calibrated against four naive baselines. Cached artefacts feed the Q1, Q2, and Q3 analyses.

Figure 1 summarises the end-to-end pipeline. The remainder of this section walks through each stage in turn.

3.1. Dataset and Problem Setup

The primary dataset is WikiChurches [1], a fine-grained architectural-style dataset of European church buildings. The annotation file defines 106 feature types (arches, windows, towers, portals, etc.) with bounding boxes in normalised coordinates. This project uses two data scopes: (1) an expert-annotated evaluation subset of 139 images with 631 bounding boxes (Romanesque 51, Gothic 49, Renaissance 22, Baroque 17), and (2) a larger style-labelled pool from the 9,485-image official release used for fine-tuning. The 139 annotated images are held out from all fine-tuning splits.

A key caveat is *sparse annotation bias*: WikiChurches annotates representative instances rather than every visible occurrence of a feature, so a model attending to multiple correct instances may still be penalised by IoU. This is treated as a documented limitation rather than a pipeline defect.

3.2. Models and Attention Extraction

The frozen benchmark compares seven models: six ViT-Base transformers (DINOv2 [3], DINOv3 [4], MAE [15], CLIP [16], SigLIP [17], SigLIP2 [18]) and ResNet-50 as a supervised CNN baseline. DINOv2 uses patch size 14 with 4 register tokens (16×16 spatial grid); all others use patch size 16 (14×14 grid).

Attention extraction depends on architecture (Table 1). CLS attention and rollout are used for DINOv2, DINOv3, MAE, and CLIP. SigLIP and SigLIP2 use a mean-attention proxy because they lack an equivalent CLS-token pathway. ResNet-50 uses Grad-CAM. These distinctions are especially important in Q3, which is restricted to architecture-native CLS-token models.

Table 1: Attention extraction methods by model family.

Method	Models	Notes
CLS attention	DINOv2, DINOv3, MAE, CLIP	Native CLS token
Rollout	DINOv2, DINOv3, MAE, CLIP	Cross-layer composition
Mean attention	SigLIP, SigLIP2	Proxy (no CLS token)
Grad-CAM	ResNet-50	CNN baseline

3.3. Alignment Metrics

Five metrics are used because no single score suffices for all interpretations. Let A denote the attention heatmap, M the union binary mask of expert boxes, and G a Gaussian soft-union target derived from the boxes.

IoU thresholds A at the exact top-k pixel count via `torch.topk` (top 10% of pixels) and computes overlap with M:

$$\text{IoU} = \frac{|\hat{A} \cap M|}{|\hat{A} \cup M|}, \quad \hat{A} = \text{top-}k\text{-pixels}(A). \tag{1}$$

Higher is better.

Coverage is threshold-free: the fraction of total attention energy inside M,

$$\text{Coverage} = \frac{\sum_{i \in M} A_i}{\sum_i A_i}. \tag{2}$$

Higher is better.

MSE, **KL divergence**, and **EMD** compare the full normalised heatmap against G. Lower is better for all three.

3.4. Baselines and Calibration

Continuous metrics are calibrated against four naive baselines: random attention, center Gaussian, saliency prior, and Sobel edge map (Table 2). Beating random only is weak evidence; clearing all four baselines is stronger support for non-trivial semantic alignment.

Table 2: Dataset-level mean baseline scores (lower is better).

Baseline	MSE	KL	EMD
Random	0.3192	3.3627	0.3468
Center Gaussian	0.1770	2.6317	0.2836
Saliency Prior	0.0957	2.6111	0.2654
Sobel Edge	0.0376	3.2237	0.3137

3.5. Fine-Tuning Protocol

Q2 uses a shared experiment-batch workflow. Training is on the non-annotated style-labelled pool with one stratified validation split shared across all model$\times$strategy runs. Best checkpoints are selected by classification validation accuracy; alignment is re-evaluated on the held-out annotated subset. Three strategies are compared (Table 3); ResNet-50 is excluded from Q2.

Table 3: Fine-tuning strategies compared in Q2.

Strategy	Backbone	Role
Linear Probe	Frozen	Control (zero attention change)
LoRA	Mostly frozen	Efficient adapter
Full	Updated end-to-end	Maximum adaptation

3.6. Per-Head Scope (Q3)

Q3 is scoped to architecture-native CLS-token models: DINOv2, DINOv3, MAE, and CLIP, with frozen, LoRA, and Full as the primary variants. Linear Probe is a control. SigLIP, SigLIP2, and ResNet-50 are excluded because per-head analysis for them relies on a proxy rather than native attention heads. Q3 is a descriptive head-specialisation analysis, not a causal attribution method.

3.7. Statistical Analysis

The study uses paired t-tests, Wilcoxon signed-rank tests, bootstrap confidence intervals, Cohen's d for paired differences, and Holm multiple-comparison correction. This is especially important in Q2, where many model$\times$strategy$\times$metric combinations are compared within shared correction families.

4. EXPERIMENTAL RESULTS

4.1. Dataset

The 139-image annotated evaluation set spans four architectural styles. Average expert boxes per image are 4.5 overall, but vary by style: Gothic and Romanesque images are more densely annotated (4.2–5.9 boxes/image) than Baroque images (1.8 boxes/image), which limits statistical power for that subset.

4.2. Implementation Details

All ViT models use ViT-Base scale. Fine-tuning used AdamW with cosine learning rate scheduling, early stopping on validation accuracy. LoRA adapters are inserted into attention projection layers. Heatmaps are computed at a standardised resolution and cached to HDF5/SQLite for fast backend serving.

4.3. Performance Metrics

Metrics are defined in Section 3. IoU@90 denotes IoU at the top-10% pixel threshold. Directions: higher is better for IoU and Coverage; lower is better for MSE, KL, and EMD. Results are reported at each model's best default-method layer over the 139 annotated images.

4.4. Q1: Frozen-Model Attention Alignment

Attention alignment is not one thing. A model can place its strongest attention inside the expert boxes, spread attention across the right facade region, or match the overall target distribution - and those are related but not identical behaviours. Treating one score as the whole answer would make the result look cleaner than it really is. Instead, Q1 uses a multi-metric benchmark: IoU@90, Coverage, MSE, KL divergence, and EMD.

Each metric catches a different failure mode: IoU@90 asks whether the model's top-attended pixels land inside the expert annotations; Coverage asks what fraction of total attention energy falls inside the annotated regions; MSE, KL, and EMD compare the full heatmap against a Gaussian soft-union target, exposing cases where a model looks reasonable under overlap but still places attention mass in the wrong location.

Under this benchmark, DINOv3 is the cleanest Q1 result (Table 4). It leads on IoU@90, Coverage, KL, and EMD, and is the only model that clears all four naive baselines on all three continuous metrics. It does not win MSE, where the extra displayed precision shows SigLIP2 marginally lower than base SigLIP (0.01745 vs. 0.01755) - exactly why the benchmark cannot collapse to one number. The SigLIP rows also show that `siglip` and `siglip2` should not be treated as interchangeable: base SigLIP reaches a competitive IoU@90 peak of 0.0739 at `layer8`, ahead of MAE and CLIP, while SigLIP2 remains weaker on sharp overlap despite its slightly lower MSE. SigLIP and SigLIP2 also pair good MSE with EMD ≈ 0.35, worse than the random baseline of 0.3468: smooth local maps can still misplace attention mass at a distributional level. CLIP's best frozen IoU occurs at layer 0 rather than in a late-layer regime, consistent with its global contrastive objective applying no patch-level spatial pressure.

Table 4: Q1 frozen-model alignment leaderboard (best default-method layer, 139 images). ↑ higher is better; ↓ lower is better.

Model	Paradigm	IoU↑	Cov↑	MSE↓	KL↓	EMD↓
DINOv3	Self-distil.	**0.133**	**0.137**	0.0270	**2.325**	**0.260**
ResNet-50	Supervised	0.090	0.104	0.0242	2.692	0.303
DINOv2	Self-distil.	0.082	0.100	0.0209	2.684	0.298
SigLIP	Sigmoid CL	0.074	0.076	0.01755	3.002	0.348
MAE	Masked recon.	0.070	0.090	0.0483	2.756	0.318
CLIP	Lang.-image	0.049	0.085	0.0211	2.912	0.326
SigLIP2	Sigmoid CL	0.047	0.071	**0.01745**	3.071	0.354

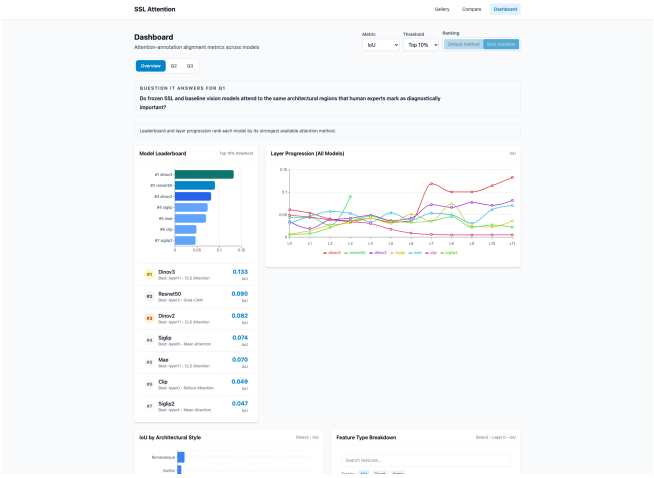


Fig. 2: React dashboard screenshot (`/dashboard`, IoU@90, Best available ranking). DINOv3 ranks first at layer 11 with IoU = 0.133 using CLS attention. The leaderboard also shows base SigLIP at rank 4 with IoU = 0.074 at `layer8`, separated from SigLIP2 at rank 7. The layer-progression panel shows DINOv3's late-layer jump relative to other models.

Gram Anchoring

Our hypothesis is that DINOv3 benefits from the combination of scale, curated data, and a training recipe designed to preserve dense spatial structure. The DINOv3 technical report [4] identifies dense-feature degradation as a failure mode during long SSL training and introduces **Gram anchoring** to stabilise patch-level feature maps. Given that Q1 metrics reward spatial correspondence to expert-marked architectural

parts, a method designed to protect dense features is a natural fit for the observed result.

To corroborate DINOv3's lead, we performed two further checks. The first is robustness: paired image-level tests (Table 5) show DINOv3 remains separated from the next-best model on all headline metrics, with Holm-adjusted $p < 1.31 \times 10^{-7}$. The second is where DINOv3 wins. If Gram anchoring preserves cleaner patch-level spatial structure, we should expect frozen attention to work best on mid-to-large architectural parts and still struggle on small ornamentation - and that is what we observe.

Table 5: Q1 robustness and spatial checks for DINOv3.

Check	Result	Why it matters
Paired gap	+0.0425 IoU, +0.0330 Cov, +0.3595 KL, +0.0378 EMD	Lead is not a leaderboard artifact
Strongest styles	Gothic 0.169, Romanesque 0.160	Highest where structure is visually prominent
Strongest features	Ornate Portal 0.214, Tracery Rose 0.164	Aligns with large coherent parts
Weakest features	Crocket, Fleuron, Pinnacle	Small ornamentation remains hard

This does not prove Gram anchoring is the cause - the current artifact does not ablate model scale, data mixture, and the anchoring loss separately - but it makes the hypothesis plausible: DINOv3 has a better dense spatial prior for prominent architectural structure, not a complete understanding of every fine-grained architectural cue.

4.5. Q2: Fine-Tuning Effects on Attention

Linear Probe produces exactly zero Δ across all metrics for all models, confirming that observed Q2 movement is tied to actual representation change rather than a reporting artefact. Figure 3 shows the multi-metric improvement heatmap across all non-Linear-Probe combinations.

CLIP shows the largest gain (Full fine-tuning: IoU $0.0181 \rightarrow 0.0745$, Coverage $0.051 \rightarrow 0.105$, Cohen's $d \approx 1.0$). The per-style breakdown (Table 6) reveals that virtually all of CLIP's gain concentrates on Romanesque (+0.066) and Gothic (+0.079) images, whose diagnostic features (Round Arch Portals, Pointed Arch Portals, Tracery) appear frequently in English-language web text - consistent with CLIP's language-grounded training. Renaissance (+0.014) and Baroque (+0.013) show near-zero gain.

A Kruskal-Wallis test finds significant style moderation for CLIP ($p = 7.2 \times 10^{-9}$), MAE ($p = 3.8 \times 10^{-7}$), SigLIP ($p = 1.5 \times 10^{-5}$), and SigLIP2 ($p = 5.0 \times 10^{-7}$); DINOv2 and DINOv3 are not significant ($p = 0.18$, $p = 0.12$), consistent with their near-zero per-style profiles.

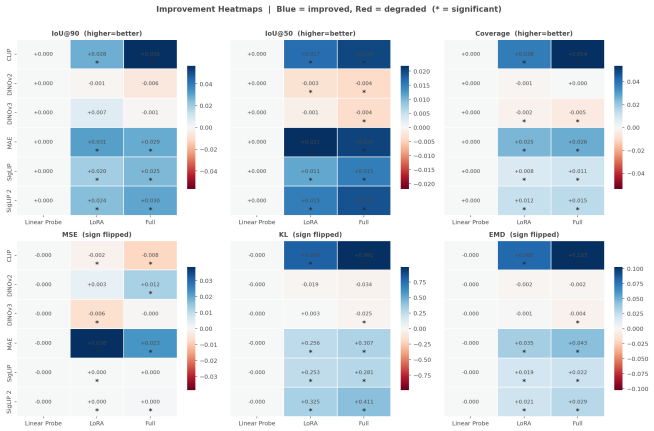


Fig. 3: Q2 multi-metric improvement heatmap. Each cell shows whether LoRA or Full fine-tuning enhances, preserves, or degrades alignment relative to the frozen model.

Table 6: Per-style ΔIoU@90 under Full fine-tuning (frozen $\rightarrow$ fine-tuned).

Model	Roman.	Gothic	Renaiss.	Baroque
CLIP	+0.066	+0.079	+0.014	+0.013
MAE	+0.007	+0.009	+0.108	+0.045
SigLIP2	+0.034	+0.044	+0.007	+0.007
SigLIP	+0.029	+0.039	−0.006	+0.005
DINOv2	−0.010	+0.001	−0.004	−0.012
DINOv3	−0.001	+0.006	−0.004	−0.009

MAE's Renaissance Δ of +0.108 is the largest single-style shift in the dataset. Table 7 shows this gain is concentrated on pediment-class features, which are geometrically compact and Renaissance-exclusive. The two most common Renaissance features (Pilaster, Belt Course) show negative Δ, suggesting the style-classification gradient routes attention toward the most discriminative geometric forms.

SigLIP variants. The Q2 analysed layer is `layer11`, where both SigLIP and SigLIP2 start from weaker frozen IoU@90 baselines than the DINO family (`siglip` 0.0364, `siglip2` 0.0220). This is not the Q1 best-layer story - base SigLIP's stronger peak sits at `layer8` - so Q2 is sharpening a relatively weak late-layer signal rather than improving each model's best frozen layer. Under Full fine-tuning, base SigLIP remains higher in absolute IoU@90 (0.0618 vs. SigLIP2 0.0519), but SigLIP2 moves further in standardised-effect terms ($\Delta = 0.0299$, Cohen's $d = 0.781$, vs. SigLIP's $\Delta = 0.0254$, $d = 0.604$), consistent with more late-layer adaptation headroom from a weaker starting point and the denser grounding components of SigLIP2.

DINO family. Both DINOv2 and DINOv3 show near-

Table 7: Per-feature ΔIoU for MAE (Full fine-tuning, Renaissance images).

Feature	Froz. IoU	FT IoU	ΔIoU
Triangular Pediment	0.036	0.116	+0.080
Cranked Cornice	0.005	0.067	+0.062
Broken Pediment	0.005	0.060	+0.055
Volute	0.009	0.052	+0.043
Segmental Pediment	0.013	0.054	+0.042
Pilaster	0.023	0.011	−0.012
Belt Course	0.031	0.015	−0.016

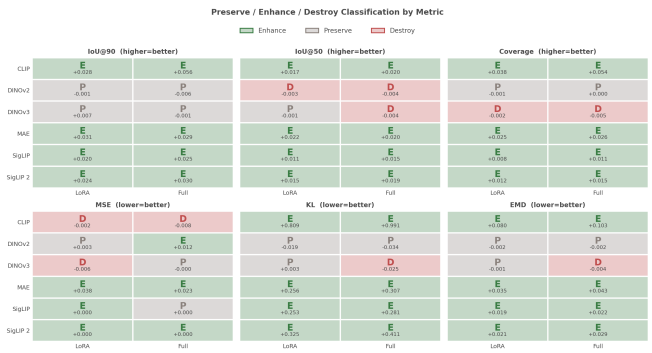


Fig. 4: Preserve/enhance/destroy classification across 72 non-LP combinations. Fine-tuning dominantly enhances alignment, with concentrated regression risk.

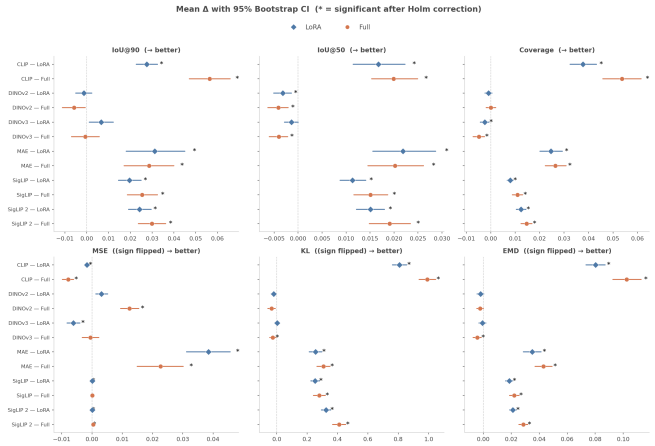


Fig. 5: Forest plot of ΔIoU@90 with 95% bootstrap confidence intervals per model and strategy.

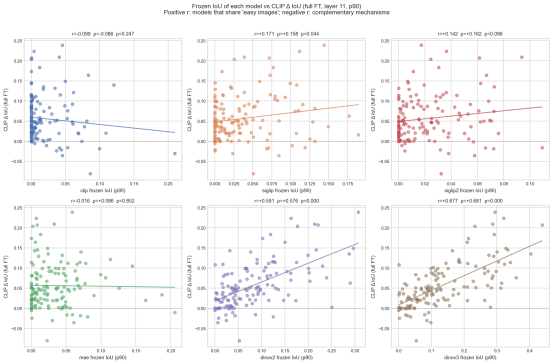


Fig. 6: DINOv3 frozen IoU@90 vs. CLIP ΔIoU@90 per image ($r = +0.677$). Structural image difficulty, not model family, determines what is "easy".

zero Δ across all styles and strategies. $\Delta \approx 0$ is a positive generalisation result: expert-aligned spatial structure is already encoded by pretraining, and the style-classification gradient is absorbed by the classification head rather than reshaping backbone attention.

Figure 4 shows the preserve/enhance/destroy classification across 72 non-Linear-Probe model×strategy×metric combinations: 46 enhance, 16 preserve, 10 destroy. Figure 5 adds the statistical layer with bootstrap confidence intervals.

Cross-model correlation. DINOv3 frozen IoU predicts CLIP ΔIoU at Pearson $r = +0.677$ ($\rho = +0.681$, $p < 0.0001$; see Figure 6): the images CLIP gains on are the same structurally easy images DINOv3 already attends to. Within-cluster language-model correlations (CLIP–SigLIP–SigLIP2) are $r \approx 0.43$–0.58; MAE is anti-correlated with the language cluster ($r \approx -0.22$ to -0.31). Models with different pretraining objectives converge on the same structurally easy images rather than specialising on complementary subsets.

4.6. Q3: Per-Head Specialisation

Q3 defines *descriptive specialisation* as a head whose CLS-to-patch attention map consistently aligns better than other heads with expert-marked regions or feature labels. It is not a causal-attribution claim. The scope is the architecture-native CLS-token ViTs: DINOv2, DINOv3, MAE, and CLIP; SigLIP, SigLIP2, and ResNet-50 are excluded because their per-head proxies do not use native attention heads, and rollout is excluded because it aggregates layers. Heads are ranked primarily by IoU@90, with Coverage and EMD as robustness checks.

Head Ranking.

Expert-aligned attention concentrates in a small number of heads rather than spreading uniformly across the transformer (Figure 7). DINOv3 is the cleanest case: the same `layer10/head8` remains best across Frozen, LoRA, and Full, appearing in the top three on more than 110 of 139 images in every condition. DINOv2 shows the same preserved-head pattern at lower absolute alignment. MAE shifts from `layer10/head5` to `layer11/head7` under LoRA, then returns to `layer10/head5` under Full. CLIP is the clearest reorganisation case: frozen CLIP's best heads sit in early

layers (`layer4/head5` leads at mean IoU@90 = 0.067), while the selected late layer used in the delta view is weak before adaptation (`layer11` max IoU@90 = 0.034) and then becomes the strongest adapted layer (best `layer11` head reaches 0.084 under LoRA and 0.105 under Full), with `layer4/head5` staying near its frozen score. Coverage and EMD agree on the high-level reading, with DINO-family stability being cross-metric while MAE and CLIP preserve only the broader adaptation pattern rather than the exact winning head.

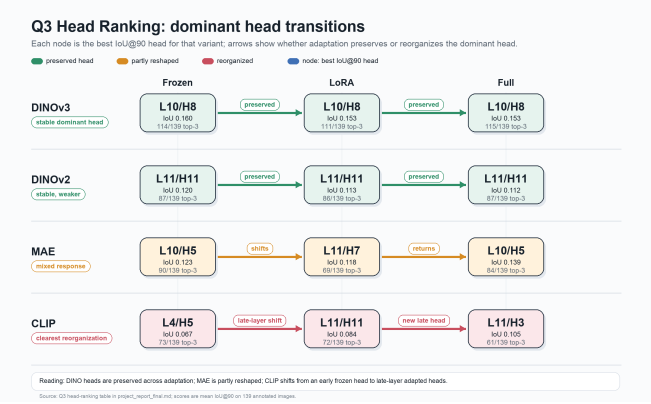


Fig. 7: Q3 head-ranking transition map. Each node is the best IoU@90 head for a model variant, with mean IoU@90 and top-3 frequency across the 139 annotated images. DINO-family models preserve their dominant heads across adaptation; MAE shows a mixed pattern; CLIP shifts from an early frozen head to late-layer adapted heads.

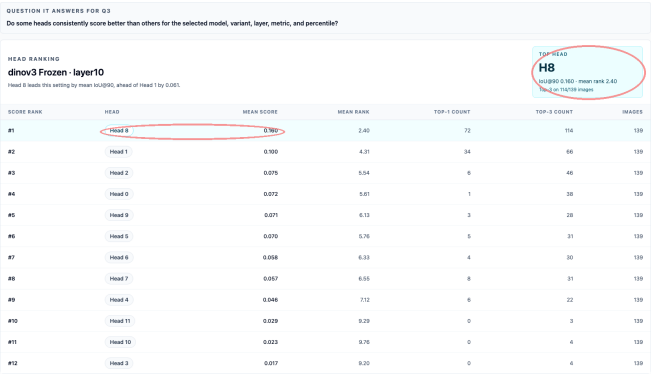


Fig. 8: DINOv3 frozen head-ranking drill-down at `layer10` using IoU@90. Head 8 leads with mean IoU@90 = 0.160, mean rank 2.40, and top-3 placement on 114/139 images, supporting the sparsity claim behind the transition map.

The same concentration pattern holds across the full 12×12 grid of (layer, head) pairs. In each frozen scoped model, the top-ranked pair's mean IoU@90 lands $2.7\times$–$3.5\times$ above the median pair's (DINOv3 $3.45\times$, CLIP $3.22\times$, MAE $2.92\times$, DINOv2 $2.69\times$). The five best pairs together account for 7.3%–9.0% of total mean IoU@90 across all 144 pairs, roughly $2\times$ the 3.5% they would hold if alignment were uniform. This is the quantitative content of "sparse": expert-aligned signal concentrates in a small minority of heads rather than spreading evenly.

This connects to Q2. DINO-family models already have strong frozen spatial alignment, so fine-tuning mostly preserves the dominant expert-aligned head. CLIP and MAE gain more from adaptation, and Q3 shows that those gains can come from *different heads* becoming best aligned after fine-tuning. CLIP's `layer11` alignment is newly strengthened by adaptation rather than inherited from an already strong frozen `layer11` head, while MAE is selectively reshaped toward discriminative geometric forms.

Head-Feature Matrix.

Head Ranking identifies a dominant head; the Head-Feature Matrix asks what architectural evidence that head aligns with. The observation is that the strongest heads mostly align with larger, coherent structural features (Figure 9). For DINOv3, the same `layer10/head8` that leads the ranking view also carries the selected *Columned Portal* cell at IoU@90 = 0.215 across 15 annotations, linking the ranking evidence to the feature evidence.

To keep the visual check disciplined, for each scoped Q3 model we take its frozen IoU@90 dominant head from the ranking view and sort architectural features for that head with at least three annotations. Figure 10 shows the three strongest and three weakest features for each model's dominant head.

Across the four scoped models the pattern is consistent. DINOv3's `layer10/head8` reaches mean feature-level IoU@90 of 0.215 on *Columned Portal*, 0.204 on *Round Arch Portal*, and 0.181 on *Ornate Portal*, while its weakest supported features are essentially zero on *Blind Tracery*, *Crocket*, and *Tabernacle*. DINOv2 shows the same portal-heavy top end and the same small-detail failure boundary. MAE also favours large facade parts, but its third-strongest feature is *Wimperg*, fitting the Q2 observation that MAE is more responsive to compact geometric forms. CLIP's inclusion of *Belt Course* suggests feature extent and clean geometry matter, not only semantic category names.

The strongest heads are therefore better described as heads whose spatial patterns are more compatible with expert-marked *structural* parts rather than as exact part detectors, and the failure cases are not random: across families, the weak features tend to be small, thin, repeated, or visually entangled with surrounding masonry.

One caveat moderates the weak-feature reading. IoU@90 keeps a fixed-size top-10% attention region and scores it against a variable-size expert mask, so for thin or repeated ornamentation (*Crocket*, *Fleuron*, *Blind Tracery*) the mask is small and the achievable IoU@90 is mechanically capped regardless of where attention falls. Part of the weak-feature

Fig. 9: Head-Feature Matrix report view for DINOv3 frozen attention (`layer = 10`, IoU@90). The same `head8` that leads the ranking view also carries the selected *Columned Portal* cell at IoU@90 = 0.215 across 15 annotations: the dominant head is strongest on portal-scale structure.

Fig. 10: Frontend Image Detail Q3 crops for the frozen dominant IoU@90 head in each scoped model: DINOv3 `layer10/head8`, DINOv2 `layer11/head11`, MAE `layer10/head5`, CLIP `layer4/head5`. Strongest features are dominated by portal-scale or facade-structure parts (*Columned Portal*, *Round Arch Portal*, *Ornate Portal*, *Wimperg*, *Belt Course*). Weakest features are thin or decorative labels (*Blind Tracery*, *Crocket*, *Tabernacle*, *Fleuron*, *Coupled Twin Window with Discharging Arch*, *Projecting Cornice*).

shortfall is therefore metric geometry rather than head-attribution; a threshold-free Coverage check would help separate the two effects.

Frozen-to-Adapted Delta.

Head Ranking shows whether the winning head changes across variants. The Frozen-to-Adapted Delta view asks the sharper follow-up: when adaptation changes the dominant head, does the strongest expert-aligned signal stay in the same head or move to a different one? Here, *expert-aligned signal* means the selected head's normalised CLS-to-patch heatmap scored against expert boxes; for IoU@90 the pipeline keeps the top 10% of heatmap pixels and compares that binary mask with the expert-box union, and the head with the highest mean IoU@90 across the 139 images is treated as the strongest expert-aligned head for that layer, variant, and metric.

The answer is family-specific: DINO-family heads are mostly preserved, MAE is partially reshaped, and CLIP shows the clearest reorganisation. The delta view makes the CLIP case concrete inside `layer11`, the late layer where adapted CLIP becomes strongest. The cross-layer change is that frozen CLIP's best alignment lives in early layers (`layer4` max IoU@90 = 0.067), while `layer11` is weak before adaptation (0.034 max). Adaptation strengthens `layer11` (0.084 LoRA, 0.105 Full) without substantially changing `layer4/head5`; this view then zooms into `layer11` to show which adapted heads pick up the new signal: frozen `H4` gives way to LoRA `H11` and Full `H3` (Figure 11).

Reading the four scoped models together: DINOv3 is the clean stability case, with `layer10/head8` remaining the strongest IoU@90 head in Frozen, LoRA, and Full and top-3 support above 110/139 images in every condition; DINOv2 shows the same preserved-head pattern at lower absolute alignment with `layer11/head11`. MAE is the intermediate case - Frozen and Full both favour `layer10/head5`, while LoRA shifts the strongest IoU@90 head to `layer11/head7`: not a full CLIP-style rewrite, but evidence that parameter-efficient adaptation can move MAE's dominant signal. CLIP is the clearest reorganisation case: in `layer11`, LoRA promotes `H11` from rank #6 to #1, with frozen-best `H4` only falling to #3 and remaining close in score (0.080 vs. 0.084), so LoRA promotes a new top head without erasing the frozen late-layer pattern. Full fine-tuning reorganises CLIP more strongly - `H3` moves from #8 to #1, `H8` moves from #7 to #2, and frozen-best `H4` drops to #4. Read with Head Ranking, this gives the CLIP-specific claim: adaptation strengthens expert alignment in `layer11` (where frozen CLIP is weak, 0.034 max) and reorganises the within-layer ranking of those newly strengthened heads.

DINO-style models already align well with expert regions

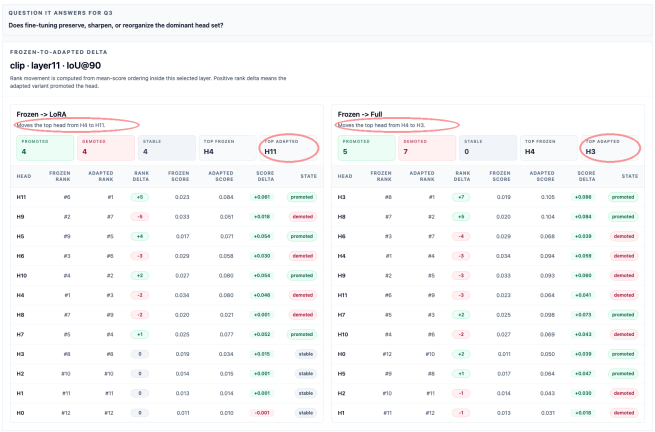


Fig. 11: Frozen-to-Adapted Delta report view for CLIP (`layer = 11`, IoU@90). Within the same late layer, the LoRA comparison moves the top head from `H4` to `H11`, while the Full comparison moves it from `H4` to `H3`.

before fine-tuning, so adaptation leaves their strongest heads mostly unchanged. CLIP and MAE have more room to move: after adaptation, the best-aligned attention comes from *different heads*, not just from the same head with a higher score.

4.7. Ablation Study: Preserve / Enhance / Destroy

The preserve/enhance/destroy classification in Figure 4 serves as the primary ablation across the full model$\times$strategy$\times$metric matrix. Key findings: (1) Linear Probe is a true zero-change control across all models; (2) Full fine-tuning produces the most enhance outcomes overall but also the most destroy outcomes for already-aligned models; (3) LoRA captures a substantial share of Full fine-tuning's improvement while producing fewer regression cases for the DINO family.

4.8. Discussion and Limitations

Intuitive findings. Linear Probe as a true attention-change control validates that Q2 movement reflects representation change, not a reporting artefact. Stronger task-conditioned adaptation helps most when the frozen model is not yet strongly aligned - CLIP, MAE, and SigLIP-family results fit this pattern.

Surprising findings. The SigLIP family pairs excellent frozen MSE with EMD worse than random, illustrating why single-metric interpretation is dangerous. DINOv2 does not improve dramatically under Q2 despite other families gaining substantially: self-distillation may already produce more part-like spatial attention, leaving less headroom for downstream adaptation.

Pretraining objective account. CLIP's InfoNCE loss [16] applies no patch-level spatial pressure; fine-tuning is the first spatial signal, and the gain concentrates on styles whose features are linguistically described. MAE's 75%-masking objective [15] forces precise local geometry representations, which explains its Renaissance pediment affinity. DINOv3's Gram-anchoring loss [4] explicitly preserves second-order spatial structure, explaining its $\Delta \approx 0$ result under Q2.

Practical implications. Model selection for domain adaptation should not be guided by accuracy alone. DINOv3 may be preferable when plausible evidence use matters without adaptation; CLIP or MAE become more attractive when adaptation is allowed. LoRA captures most of the attention-shift benefit in several models with fewer regressions than Full fine-tuning. MAE is the only model whose per-image Δ is anti-correlated with the language cluster, making it the natural complementary partner in coverage-seeking ensembles.

Limitations. The annotated evaluation subset is small (139 images), limiting statistical power. Bounding boxes are expert-guided but not exhaustive, introducing sparse annotation bias especially for Baroque. Attention is an incomplete explanation signal: strong alignment is evidence of plausibility, not causal proof. IoU depends on the thresholding rule; continuous metrics require calibration to avoid overreading raw numbers.

5. CONCLUSIONS AND FUTURE WORK

This project asks whether self-supervised vision models attend to the same architectural evidence that experts consider diagnostically important. Using WikiChurches, expert bounding boxes, and a multi-metric alignment framework, we address three linked questions across seven models and three fine-tuning strategies.

For **Q1**, frozen expert-aligned attention exists but is not evenly distributed. DINOv3 provides the strongest evidence, leading the default-method benchmark on IoU, Coverage, KL, and EMD and clearing all calibrated baselines. Base SigLIP is a stronger frozen overlap result than SigLIP2 at its best mean-attention layer, but both variants still have late-layer adaptation headroom. The safest interpretation is not that all SSL models attend like experts, but that DINOv3's spatial prior is unusually compatible with expert-marked architectural evidence in WikiChurches.

For **Q2**, fine-tuning's effect is mediated by three interacting factors: the pretraining objective's spatial prior (DINO preserves strong frozen alignment; CLIP, MAE, and SigLIP-family gain because they lack it); dataset linguistic coverage (CLIP's gain concentrates on Gothic and Romanesque features densely described in English web text); and geometric discriminability (MAE redirects attention to Renaissance-exclusive pediment forms that pixel reconstruction already encodes with high local fidelity). Models with different objectives converge on the same structurally easy images rather than complementary subsets, with implications for ensembling strategies.

For **Q3**, per-head specialisation is sparse, descriptive, and family-shaped. DINOv3's `layer10/head8` and DINOv2's

`layer11/head11` remain dominant across Frozen, LoRA, and Full, consistent with the DINO family's already-strong spatial prior. MAE is partly reshaped by adaptation, while CLIP shows the clearest reorganisation from an earlier frozen head to later adapted heads. The strongest head-linked architectural features cluster on portals, arches, and rose-window-scale structures, so the safest claim is not that heads are exact architectural-part detectors, but that some heads expose spatial patterns more compatible with expert-marked structure than others.

Future work. A patch-level text-similarity probe for CLIP would directly test the linguistic-grounding hypothesis. Extending the annotated pool beyond 139 images would improve statistical power and reduce Baroque annotation sparsity. A Gram-anchoring ablation on DINOv3 would isolate whether the spatial prior is driven by data scale, model scale, or the anchoring loss specifically.

6. AUTHOR CONTRIBUTIONS

Desmond Choy: project lead, analysis interface, precompute pipeline, metric design and validation, Q1 calibration, Q3 analysis and backend.

Leong Kay Mei: frontend and analysis interface, fine-tuning pipeline, Q2 analysis, result interpretation, and documentation.

Both authors contributed to research design, experiment execution, and report writing.

7. AI TOOL DECLARATION

The authors used Claude (Anthropic) to assist with literature review summarisation, LaTeX formatting and table layout, drafting prose for subsequent revision, and code-comment polishing in the analysis pipeline. All experimental design, results, and interpretations are the authors' own, and the authors are responsible for the content and quality of the submitted work.

8. REFERENCES

[1] Björn Barz and Joachim Denzler, "WikiChurches: A fine-grained dataset of architectural styles with real-world challenges," 2021.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin, "Emerging properties in self-supervised vision transformers," in *IEEE/CVF Int. Conf. on Computer Vision*, 2021, pp. 9650–9660.

[3] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski, "DINOv2: Learning robust visual features without supervision," *Trans. on Machine Learning Research*, 2024.

[4] Oriane Siméoni, Timothée Darcet, Théo Moutakanni, Mathilde Caron, Hugo Touvron, Nicolas Ballas, Piotr Bojanowski, Armand Joulin, and Maxime Oquab, "DINOv3: Scaling self-supervised learning," 2025.

[5] Seungwon Chung et al., "What should we learn from attention maps? A ViT study in medical imaging," 2025.

[6] Samira Abnar and Willem Zuidema, "Quantifying attention flow in transformers," in *Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4190–4197.

[7] Hila Chefer, Shir Gur, and Lior Wolf, "Transformer interpretability beyond attention visualization," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2021, pp. 782–791.

[8] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," in *Int. Conf. on Learning Representations*, 2022.

[9] Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham, "LoRA learns less and forgets less," 2024.

[10] Alexander C. Li et al., "On the surprising effectiveness of attention transfer for vision transformers," 2024.

[11] Matthew Walmer, Soheil Suri, Kartik Gupta, and Abhinav Shrivastava, "Teaching matters: Investigating the role of supervision in vision transformers," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2023, pp. 7486–7496.

[12] Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun, "What do self-supervised vision transformers learn?," in *Int. Conf. on Learning Representations*, 2023.

[13] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov, "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned," in *Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 5797–5808.

[14] Sheng Li et al., "Interpreting vision transformer from head distribution," *IEEE Trans. on Visualization and Computer Graphics*, 2023.

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick, "Masked autoencoders are scalable vision learners," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2022, pp. 16000–16009.

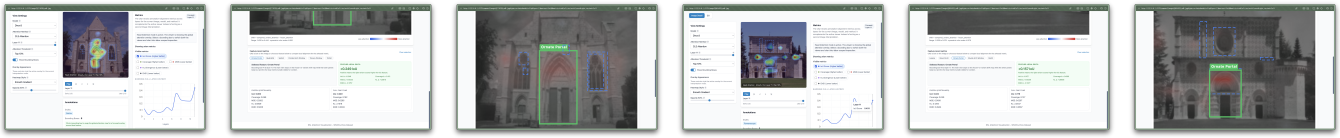
[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision," in *Int. Conf. on Machine Learning*, 2021, pp. 8748–8763.

[17] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer, "Sigmoid loss for language image pre-training," in *IEEE/CVF Int. Conf. on Computer Vision*, 2023, pp. 11975–11986.

[18] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Alaaeldin El-Nouby, Xiaohua Zhai, Basil Mustafa, and Lucas Beyer, "SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," 2025.

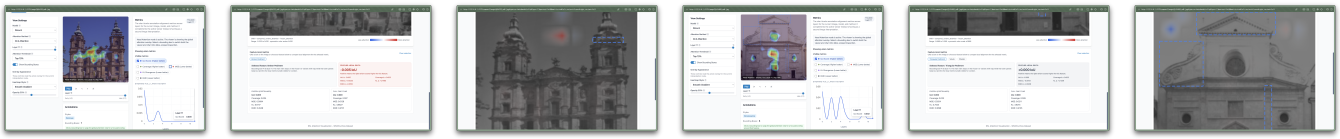
Appendix A: Structurally Easy vs. Hard Images

Each row shows one representative church from the annotated evaluation set. Left: DINOv3 frozen CLS attention at layer 11 with expert boxes overlaid. Centre: CLIP frozen-vs-fine-tuned shift map (blue = gained, red = lost attention). Right: shift map with selected expert feature box highlighted.



DINOv3 frozen | CLIP shift | Feature box | DINOv3 frozen | CLIP shift | Feature box

Easy images (high DINOv3 frozen IoU, high CLIP ΔIoU). Q1710328 Gothic: DINOv3 IoU $= 0.438$; CLIP ΔIoU $= +0.207$ (Ornate Portal). Q2034923 Romanesque: DINOv3 IoU $= 0.403$; CLIP ΔIoU $= +0.135$ (Ornate Portal).



DINOv3 frozen | CLIP shift | Feature box | DINOv3 frozen | CLIP shift | Feature box

Hard images (low DINOv3 frozen IoU, low CLIP ΔIoU). Q694252 Baroque: DINOv3 IoU $= 0.000$; CLIP ΔIoU $= -0.005$ (Broken Pediment, peripheral). Q1424095 Renaissance: DINOv3 IoU $= 0.002$; CLIP ΔIoU $= +0.002$ (spatially diffuse annotations).

Fig. 12: Appendix A: Structurally easy (top) and hard (bottom) church images. The $r = +0.677$ correlation between DINOv3 frozen IoU and CLIP ΔIoU shows that image structure, not model family, determines alignment difficulty.