REPRESENTATIONAL ALIGNMENT ACROSS MODEL LAYERS AND BRAIN REGIONS WITH HIERARCHICAL OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard representational similarity methods align each layer of a network to its best match in another independently, producing asymmetric results, lacking a global alignment score, and struggling with networks of different depths. These limitations arise from ignoring global activation structure and restricting mappings to rigid one-to-one layer correspondences. We propose Hierarchical Optimal Transport (HOT), a unified framework that jointly infers soft, globally consistent layer-to-layer couplings and neuron-level transport plans. HOT allows source neurons to distribute mass across multiple target layers while minimizing total transport cost under marginal constraints. This yields both a single alignment score for the entire network comparison and a soft transport plan that naturally handles depth mismatches through mass distribution. We evaluate HOT on vision models, large language models, and human visual cortex recordings. Across all domains, HOT matches or surpasses standard pairwise matching in alignment quality. Moreover, it reveals smooth, fine-grained hierarchical correspondences: early layers map to early layers, deeper layers maintain relative positions, and depth mismatches are resolved by distributing representations across multiple layers. These structured patterns emerge naturally from global optimization without being imposed, yet are absent in greedy layer-wise methods. HOT thus enables richer, more interpretable comparisons between representations, particularly when networks differ in architecture or depth.

1 INTRODUCTION

Understanding high-dimensional neural activity is a shared challenge in neuroscience and artificial intelligence (AI). In neuroscience, comparing neural responses across individuals reveals which computations are universally shared versus idiosyncratic. In AI, comparing representations across models reveals how architectural choices, training objectives, and learning dynamics shape learned features, and helps identify principles of universality i.e. representational properties that emerge consistently across diverse network architectures and objectives. Comparing models to brains extends this logic further: while we cannot rerun biological evolution, we can simulate "evolution in silico" by training artificial networks with different constraints, inputs, and objectives. When such models converge on brain-like representations, they offer mechanistic hypotheses for why the brain may have adopted its computational strategies which is a deeply important theoretical question. These comparisons have revealed striking similarities between biological and artificial networks (Yamins et al. (2014); Eickenberg et al. (2017); Güçlü and Van Gerven (2015); Cichy et al. (2016); Khaligh-Razavi and Kriegeskorte (2014); Schrimpf et al. (2018; 2020); Storrs et al. (2021); Kell et al. (2018)), common computational motifs across diverse architectures and objectives Huh et al. (2024); Kornblith et al. (2019); Bansal et al. (2021); Dravid et al. (2023), and other universal representational dimensions Chen and Bonner (2025); Hosseini et al. (2024).

The standard approach to representational comparison is layer-wise matching: each source layer is paired with the single best-matched target layer under some similarity measure (e.g., Representational Similarity Analysis (Kriegeskorte et al., 2008), Centered Kernel Alignment (Kornblith et al., 2019), Procrustes distance (Williams et al., 2021) or linear predictivity). Despite its widespread use, this approach has fundamental limitations. It enforces rigid one-to-one correspondences that fail when

networks differ in depth or when a source layer corresponds to features distributed across multiple target layers. It produces asymmetric layer mappings depending on the direction of comparison and yields no unified score for global network alignment. Most importantly, by optimizing each match independently, it ignores the global activation structure and risks overfitting to noise.

We propose Hierarchical Optimal Transport (HOT), a framework for globally consistent representational alignment. Optimal transport–based methods, such as Soft-Matching distance (Khosla and Williams, 2024; Khosla et al., 2024), have recently emerged as powerful metrics for comparing neural representations. Unlike metrics such as RSA, CKA, or linear predictivity—which are rotation-invariant and thus unable to capture similarities in neuron-level tuning—OT-based methods are rotation-sensitive. They explicitly match neurons based on their tuning profiles and, by relaxing hard permutation constraints into fractional couplings, can also handle layers of unequal size. This enables richer, more flexible neuron-level alignments than either rotation-invariant similarity metrics or strict permutation-based approaches. However, Soft Matching also remains limited to pairwise layer comparisons like other methods and does not capture global structure across networks. HOT fills this gap by operating hierarchically: it simultaneously infers soft neuron-to-neuron couplings within layers and a soft, globally consistent layer-to-layer coupling across the hierarchy. Rather than forcing each source layer to match exactly one target layer, HOT allows source layers to distribute their representational "mass" across multiple target layers while minimizing the total transport cost under marginal constraints; that is, each source layer must distribute exactly 100% of its mass across target layers (no information is lost), and the total mass each target layer receives from all source layers must sum to a balanced allocation (no target is over- or under-utilized). These conservation laws ensure a balanced alignment where every layer contributes meaningfully to the global correspondence, preventing any layer from being arbitrarily overweighted or ignored in the global matching. The result is a single network-level alignment score and a soft transport plan that naturally handles depth mismatches.

We evaluate HOT on three diverse domains: comparisons between foundation models in vision (Vision Transformers like DINOv2 and ViT-MAE), large language models of varying scales (LLaMA, Qwen), and fMRI recordings from human visual cortex across different participants. Our key contributions are:

- **A principled global alignment framework (theoretical):** HOT jointly optimizes all layer correspondences to produce symmetric, globally consistent assignments with a single alignment score. In contrast, greedy pairwise approaches are asymmetric, can overweight certain layers while completely ignoring others, and and may spuriously treat wide layers as similar to every layer they are compared against, since their high dimensionality allows them to fit or partially overlap with many different representational subspaces, obscuring more meaningful correspondences.
- **Natural handling of depth mismatches (theoretical):** By allowing soft, many-to-many mappings between layers, HOT aligns networks of different depths without forcing inappropriate one-to-one correspondences.
- **Rotation-invariant extension (theoretical):** We propose an extension of HOT that incorporates additional orthogonal transformations, making the framework rotation-invariant. This ensures that correspondences can be recovered even when shared representational features are embedded in rotated subspaces, and yields consistently high-quality alignments.
- **Improved alignment scores (empirical):** Across domains (vision models, large language models, and brain data), HOT matches or surpasses standard pairwise methods, yielding higher alignment.
- **Emergent hierarchical structure (empirical):** Without imposing ordering constraints, HOT recovers known hierarchical organization in visual cortex data across subjects and reveals clean layer-to-layer correspondences in model–model comparisons where early layers map to early layers and deeper layers preserve their relative ordering. By contrast, greedy pairwise methods fail to reveal hierarchical structure, often leaving many layers unmatched while a single or few layers dominate the mappings.
- **Featural distribution across depth (empirical):** HOT reveals how deeper networks spread computations across multiple layers that shallower networks compress into fewer stages. Concretely, a single layer in a shallower network often distributes its mass across several

neighboring layers in a deeper network, an effect that greedy pairwise methods completely miss.

2 METHODS

2.1 PROBLEM SETUP AND EXISTING APPROACHES

Comparing the internal representations of neural networks often proceeds by a *pairwise* layer search under some similarity or distance measure. In general, let T denote the number of stimuli (e.g., images, text sequences) used to probe both models. We consider two neural networks:

- The first network has L layers. Layer ℓ contains n_ℓ units, and its activations across the T stimuli are represented by
$$X_\ell \in \mathbb{R}^{T \times n_\ell}, \quad \ell = 1, \ldots, L.$$
- The second network has M layers. Layer m contains n_m units, with activations
$$Y_m \in \mathbb{R}^{T \times n_m}, \quad m = 1, \ldots, M.$$

Each row of X_ℓ or Y_m corresponds to the response of all units in that layer to a single stimulus, while each column corresponds to the activity of a single unit across stimuli. For any chosen alignment metric $S(\cdot, \cdot)$ (e.g., linear predictivity, Procrustes distance, RSA, CKA, or Soft Matching), one typically does:

$$m^*(\ell) = \arg\max_m S(X_\ell, Y_m)$$

and then reports the layer-wise score $S(X_\ell, Y_{m^*(\ell)})$. This enforces a hard one-to-one mapping from each source layer ℓ to a single target layer $m^*(\ell)$. However, when $L \neq M$, or when the set of features represented in one layer of network A is distributed across multiple layers of network B, such a rigid per-layer pairing is not well-suited: a source layer may genuinely correspond to a mixture of multiple target layers. Moreover, by optimizing each layer independently, this approach ignores the *global* structure of all activations and can overfit to noise in any single layer's responses.

In what follows, we introduce a hierarchical optimal transport framework that overcomes both issues by simultaneously inferring a soft layer-to-layer coupling and per-neuron transport plans, yielding a globally balanced alignment across networks of arbitrary depths (Figure 1).

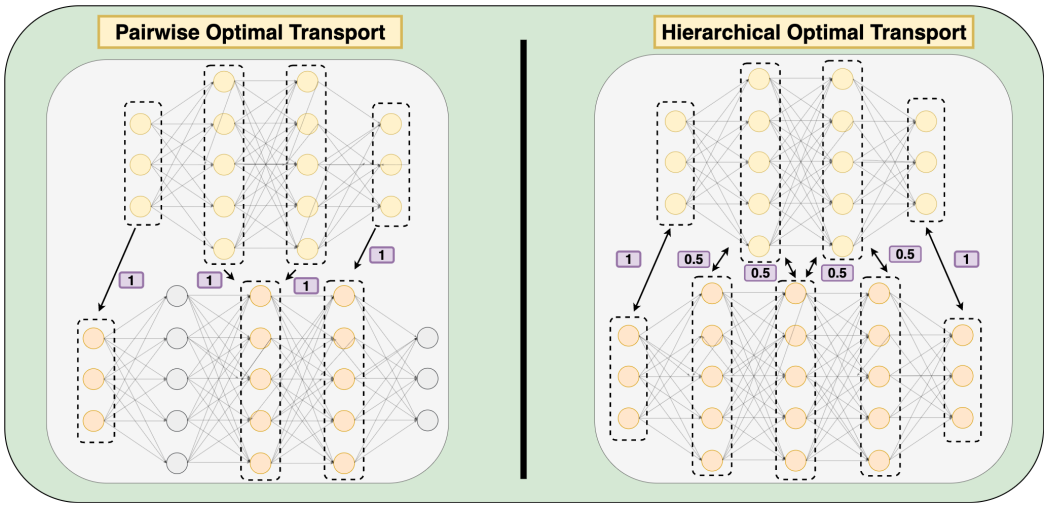

Figure 1: *Left*: **Pairwise OT**. Layers are matched independently, so multiple target layers can be mapped to the same source while other sources remain unused, yielding asymmetric, unbalanced mappings. *Right*: **Hierarchical OT**. HOT infers a globally consistent transport plan where each source layer distributes all its mass and each target layer receives exactly one unit, ensuring balanced, symmetric alignments that handle depth mismatches and reveal hierarchy.

2.2 Hierarchical Optimal Transport

Our framework operates at two hierarchical levels: an inner level that aligns neurons within each layer pair, and an outer level that determines how layers should be coupled globally.

2.2.1 Inner Level: Neuron-to-Neuron Transport

For each pair of layers (ℓ, m), we compute a cost matrix

$$C_{\ell m}^{\text{inner}}[i,j] = d_{\text{corr}}(X_\ell[:,i], Y_m[:,j])$$

where $d_{\text{corr}}(x,y) = 1 - \rho(x,y)$ is the correlation distance and ρ denotes Pearson correlation. We then solve an optimal transport problem to find a coupling $Q_{\ell m}$ that minimizes the transport cost between neurons:

$$C_{\ell m} = \min_{Q_{\ell m} \in \mathcal{T}(n_\ell, n'_m)} \langle C_{\ell m}^{\text{inner}}, Q_{\ell m} \rangle$$

where $\mathcal{T}(n_\ell, n'_m)$ is the transportation polytope:

$$\mathcal{T}(n_\ell, n'_m) = \left\{ Q \in \mathbb{R}^{n_\ell \times n'_m} : \sum_i Q_{ij} = \tfrac{1}{n'_m},\ \sum_j Q_{ij} = \tfrac{1}{n_\ell},\ Q_{ij} \geq 0 \right\}.$$

This constraint ensures that each source neuron distributes exactly $1/n_\ell$ of its mass (summing rows), and each target neuron receives exactly $1/n'_m$ of the total mass (summing columns), with all entries non-negative. The coupling $Q_{\ell m}$ specifies how neurons in layer m can be combined to reconstruct neurons in layer ℓ.

This inner-level optimization is precisely the Soft Matching distance (Khosla and Williams, 2024) applied between layers ℓ and m: it is rotation-sensitive yet permutation-invariant, and directly measures *single-neuron tuning* alignment. When the layers have equal size ($n_\ell = n'_m$), the optimal solution lies at a vertex of the transportation polytope; which is a permutation matrix by the Birkhoff–von Neumann theorem (von Neumann, 1953; De Loera and Kim, 2013). However, when layer sizes differ, soft matching assigns each neuron in one layer to weighted combinations of neurons in the other via *soft-assignments*.

2.2.2 Outer Level: Layer-to-Layer Transport

The inner costs $C_{\ell m}$ form a layer-to-layer cost matrix $C \in \mathbb{R}^{L \times M}$. We solve a second optimal transport problem to find the global layer coupling:

$$P = \arg \min_{P \in \mathcal{T}(L,M)} \langle C, P \rangle$$

where $\mathcal{T}(L,M)$ is the transportation polytope for layers:

$$\mathcal{T}(L,M) = \left\{ P \in \mathbb{R}^{L \times M} : \sum_\ell P_{\ell m} = \tfrac{1}{M},\ \sum_m P_{\ell m} = \tfrac{1}{L},\ P_{\ell m} \geq 0 \right\}.$$

The element $P_{\ell m}$ represents the fraction of layer ℓ's representation that is explained by layer m. These marginal constraints ensure mass conservation: each source layer distributes its unit mass across target layers (after normalization by L), and the total mass received by all target layers is balanced.

When the two networks have the same number of layers ($L = M$), we can show that the optimal solution P is a (scaled) permutation matrix: the objective $\langle C, P \rangle$ is linear in P, and the constraints define the transportation polytope whose vertices are permutation matrices (scaled by $1/L$) by the Birkhoff–von Neumann theorem. Since linear programs achieve their optima at vertices, the solution finds a one-to-one layer matching when $L = M$. However, when $L \neq M$, the soft coupling allows source layers to distribute their mass across multiple target layers, naturally handling depth mismatches.

2.2.3 RECONSTRUCTION AND EVALUATION

Given the layer coupling P and neuron couplings $\{Q_{\ell m}\}$, we reconstruct layer ℓ as:

$$\hat{X}_\ell = L \sum_{m=1}^{M} P_{\ell m} Y_m Q_{\ell m}^\top.$$

The factor L appears because $P_{\ell m}$ sums to $1/L$ across m. We evaluate alignment quality using the mean correlation between original and reconstructed neurons on held-out data:

$$\text{Score}_\ell = \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \rho(X_\ell[:, i], \hat{X}_\ell[:, i]).$$

The global HOT score is the average across all layers:

$$\text{HOT} = \frac{1}{L} \sum_{\ell} \text{Score}_\ell.$$

2.3 ROTATION-INVARIANT EXTENSION

Most existing metrics of representational similarity, such as RSA, CKA, and Procrustes distance, are designed to be rotation-invariant. The rationale is that when comparing population codes, we often care less about the tuning of individual neurons and more about the geometry of the representational space: distances, angles, and relative positions between stimulus responses. Two networks can encode essentially the same geometry while using different coordinate bases (for example, rotated versions of one another). Requiring strict unit-to-unit correspondence in such cases would artificially inflate dissimilarity, even though the representational geometry and information content are preserved. A rotation-invariant extension of HOT (**HOT + R** henceforth) therefore provides a way to capture this geometric equivalence while still enforcing a globally consistent layer- and neuron-level coupling. We introduce rotation matrices $R_{\ell m} \in O(n_\ell)$ for each layer pair and minimize the reconstruction error:

$$C_{\ell m} = \min_{Q_{\ell m}, R_{\ell m}} \|X_\ell R_{\ell m} - Y_m Q_{\ell m}^\top\|_F^2.$$

We optimize via alternating minimization:

- **Fix R, update Q:** Solve optimal transport using correlation distance on rotated features $X_\ell R_{\ell m}$.
- **Fix Q, update R:** Solve orthogonal Procrustes via SVD of $X_\ell^\top (Y_m Q_{\ell m}^\top)$.
- **Update P:** Refresh the outer coupling using the Frobenius reconstruction costs.

Predictions incorporate the learned rotations:

$$\hat{X}_\ell = L \sum_{m=1}^{M} P_{\ell m} Y_m Q_{\ell m}^\top R_{\ell m}^\top.$$

2.4 BASELINE COMPARISONS

We compare HOT against several baselines to isolate the contribution of each component:

Random Layer Assignment (Perm-P): We randomly permute the rows of P, breaking the optimized layer correspondences while preserving the neuron-level optimal transport within each layer pair. This control is expected to perform reasonably well because it still finds optimal neuron matches for each (now randomly assigned) layer pairing—it only disrupts which layers are matched, not how well neurons align within those matches.

HOT Top-1 Layer Transport Plan (Single-Best OT): Each source layer ℓ maps only to its highest-weight target layer

$$m^*(\ell) = \arg\max_m P_{\ell m},$$

converting the soft layer coupling to a hard one-to-one assignment while keeping the soft neuron-level transport.

Independent Pairwise OT (Pairwise Best OT): The standard greedy approach where each source layer is matched to the single target layer with minimum inner OT cost, computed independently on training data. Formally,

$$m^*(\ell) = \arg\min_m C_{\ell m},$$

and layer ℓ is reconstructed using only $Y_{m^*(\ell)}$. This baseline ignores global structure and can result in multiple source layers matching to the same target layer while leaving others unmatched.

Rotation-aware variants: We evaluate rotation-invariant versions of the pairwise OT baseline (henceforth, 'Pairwise Best + R'), where predictions use $Y_m Q_{\ell m}^\top R_{\ell m}^\top$ with orthogonal transformations optimized via Procrustes alignment.

3 RESULTS

We evaluate representational similarity under the HOT metric across four distinct alignment setups: (i) large language models of different families and scales, (ii) fMRI responses from the visual cortex of four human subjects, (iii) pretrained transformer-based vision models spanning different families and scales, and (iv) cross-domain comparisons between human visual cortex and vision transformers (Appendix Section D). Across all settings, HOT matches or surpasses baseline reconstruction (prediction) scores. Importantly, the transport plans inferred by HOT naturally reveal systematic layer-wise correspondences across models and cortex, despite not being explicitly optimized for such structure; specifically, earlier layers or regions tend to align with earlier counterparts, while deeper layers or regions map to progressively higher levels; patterns that greedy pairwise methods fail to capture.

3.1 REPRESENTATIONAL SIMILARITY BETWEEN LARGE LANGUAGE MODELS

Experimental Setup. We extract layer-wise representations by averaging token activations across 2,552 prompts from the STSB dataset (May, 2021; Enevoldsen et al., 2025; Muennighoff et al., 2022). For each model, this yields a sequence of representation matrices X and Y, corresponding to successive layers. We evaluate models of varying sizes from the LLaMA-3.2 (Grattafiori et al., 2024) and Qwen-2.5 (Qwen et al., 2025) families. Representations are compared using HOT and baseline metrics, and the resulting transport plans are analyzed. To quantify alignment quality, we reconstruct representations on a held-out validation split (20%) using the learned transport maps and report the correlation with ground-truth activations, as detailed in the Methods Section2.2.3.

Model 1	Model 2	HOT Metric	Random (Perm-P)	Single-Best OT	Pairwise Best OT
Llama-3.2 1B	Llama-3.2 3B	**0.558**	0.510	0.502	0.505
Qwen-2.5 0.5B	Qwen-2.5 3B	**0.510**	0.494	0.467	0.477
Qwen-2.5 0.5B	Llama-3.2 1B	**0.522**	0.500	0.502	0.511
Qwen-2.5 0.5B	Llama-3.2 3B	**0.531**	0.513	0.498	0.524
Llama-3.2 1B	Qwen-2.5 3B	**0.432**	0.411	0.345	0.380
Llama-3.2 3B	Qwen-2.5 3B	**0.383**	0.374	0.338	0.346

Table 1: **LLM alignment performance**. Comparison of HOT against baseline metrics, evaluated by reconstruction correlation on held-out data.

Results. HOT consistently achieves higher reconstruction accuracy than baseline methods. As shown in Table 1, reconstruction scores on the held-out validation set are significantly higher under HOT, indicating that the alignment plans it learns are more robust and generalizable than those obtained from pairwise baselines. This improvement stems from HOT's (i) enforcement of global consistency across all layers and (ii) ability to distribute representational mass across multiple target layers, allowing it to recover alignments that remain hidden to pairwise methods when features from a single layer are distributed across several layers in another model. Beyond quantitative gains, HOT uncovers clear hierarchical correspondences between models. As illustrated in Figures 2 and A.1, the transport plans produced by HOT exhibit strong diagonal structure: early layers in one model align with early layers in the other, while deeper layers align with deeper layers. Such structured correspondences are absent in pairwise OT, which often yields noisy mappings. Furthermore, when comparing shallower

with deeper models, HOT reveals that single layers in the shallower model distribute their mass across multiple consecutive layers in the deeper model. This soft many-to-many mapping reflects how additional depth refines and spreads computations, suggesting that representational stages compressed into one layer in a shallower model are decomposed across multiple processing steps in a deeper model—an organizational principle that greedy baselines fail to uncover.

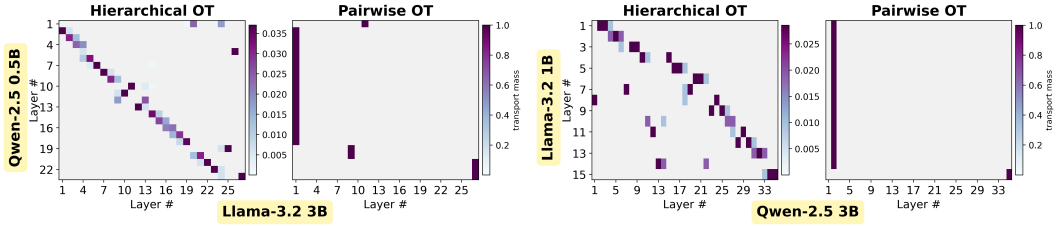

Figure 2: **Transport plans for LLM alignment.** Hierarchical OT (left) versus pairwise OT (right) for two cross-model comparisons: (a) Qwen-2.5 0.5B $\leftrightarrow$ LLaMA-3.2 3B and (b) LLaMA-3.2 1B $\leftrightarrow$ Qwen-2.5 3B. HOT reveals smooth, diagonal correspondences across layers, while pairwise OT produces noisier and less structured mappings.

3.2 Representational Similarity Between Visual Cortex across Subjects

Experimental Setup. We analyze fMRI responses from the Natural Scenes Dataset (NSD; (Allen et al., 2022)), which contains recordings from 8 participants who each viewed up to 10,000 natural images. Of these, 4 subjects viewed the full set of 10,000 images three times, with 1,000 images shared across all participants. We focus on these 4 subjects and restrict our analysis to their responses to the shared 1,000 images, ensuring that alignment is evaluated on common stimuli.

We target visual cortex regions (V1–V4), treating each region as a "layer" and individual voxels as "neurons". The visual cortex provides a strong testbed for hierarchical alignment because it is one of the best-characterized cortical systems: early areas (V1, V2) are known to encode low-level features such as orientation and contrast, while higher areas (V3, V4) encode progressively more complex shapes and object features (Hubel and Wiesel, 1968; Pasupathy and Connor, 2001; Desimone and Schein, 1987; Desimone et al., 1984). This hierarchical progression is well-established across individuals, so correspondences between homologous regions are strongly expected. HOT is applied to align responses across subjects by generating transport maps between regions. To evaluate alignment quality, we reconstruct held-out responses (20% validation split) and report correlations with ground-truth activity. We repeat this evaluation across 5 random train-validation splits to ensure robustness.

Model 1	Model 2	HOT Metric	Random (Perm-P)	Single-Best OT	Pairwise Best OT
Subject A	Subject B	0.306	0.258	0.306	**0.308**
Subject A	Subject C	0.227	0.184	0.227	**0.232**
Subject A	Subject D	0.248	0.212	0.248	**0.253**
Subject B	Subject C	**0.237**	0.199	**0.237**	**0.237**
Subject C	Subject D	0.235	0.196	0.235	**0.239**
Subject B	Subject D	0.248	0.217	0.248	**0.254**

Table 2: **Visual cortex alignment performance.** Comparison of HOT against baseline methods, evaluated by reconstruction correlation on held-out fMRI responses.

Results. As shown in Tables 2 and B.1, HOT achieves reconstruction scores comparable to pairwise OT, with only a marginal decrease in correlation. More importantly, Figures 3 and B.1 show that the transport maps inferred by HOT recover the expected cross-subject correspondences: cortical regions in one subject consistently align to the same regions in another. In contrast, pairwise OT does not produce such structured mappings in any subject pair. This indicates that HOT's global alignment scheme captures region-to-region correspondences that pairwise methods miss. To rule out the possibility that this effect stems from the OT optimization framework itself rather than from the layer-

wise setup, we additionally perform pairwise linear predictivity optimization. As shown in Figure B.2, the best-matching layers identified by linear predictivity also do not align corresponding regions, further confirming that the hierarchical alignment mechanism in HOT is critical for recovering robust and generalizable correspondences. Overall, these results demonstrate that HOT yields transport plans that are both interpretable and biologically meaningful, while maintaining reconstruction performance on par with baseline methods.

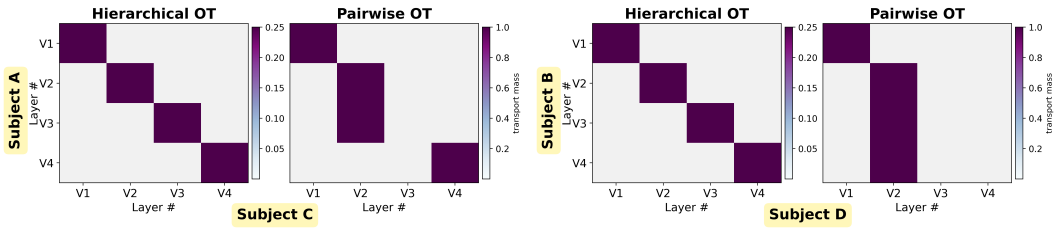

Figure 3: **Transport plans for cross-subject brain alignment.** Hierarchical OT (left) versus pairwise OT (right) for two randomly selected subject pairs: (a) Subject A $\leftrightarrow$ Subject C and (b) Subject B $\leftrightarrow$ Subject D. HOT recovers structured region-to-region correspondences that are absent in pairwise OT. Other subject pairs show similar trends (see Appendix B.1).

3.3 Representation Similarity between Vision Models

Experimental Setup. We extract layer-wise representations from Vision Transformers by averaging patch activations for each input image. We use 20,000 images randomly sampled from the ImageNet validation set (Deng et al., 2009; Russakovsky et al., 2015), sampled to ensure a uniform coverage across all classes in Imagenet. For each model, this yields a sequence of representation matrices that serve as inputs to the HOT framework. We evaluate two families of pretrained vision transformers—DINOv2 and ViT-MAE across multiple model scales. Alignment quality is quantified by reconstructing representations on a held-out validation split (20%) using the learned transport plans, with reconstruction–ground truth correlation as the metric.

Prior work has shown that the residual stream in Transformers lacks privileged axes and is invariant up to rotations of its basis (Khosla et al., 2024). Since HOT, like other OT-based methods, is rotation-sensitive, we additionally evaluate a rotation-augmented variant (HOT+R) and its baselines (see Methods Section 2.3). We restrict this analysis to Vision Transformers, where rotational invariances are especially relevant and where the computational cost of HOT+R remains tractable. For large language models and fMRI data, the added optimization over rotations was computationally prohibitive, so we report only the rotation-sensitive results in those domains. In HOT+R, the learned rotation matrices are incorporated into both transport optimization and evaluation, ensuring that geometric equivalences induced by rotations are properly captured.

Model 1	Model 2	HOT Metric	Pairwise Best OT	HOT + R	Pairwise Best + R
DINOv2 Small	ViT-MAE Base	0.289	0.301	**0.600**	0.526
DINOv2 Small	DINOv2 Large	0.353	0.340	**0.778**	0.394
DINOv2 Small	DINOv2 Giant	0.466	0.433	**0.790**	0.418
DINOv2 Small	ViT-MAE Large	0.381	0.354	**0.633**	0.509
DINOv2 Small	ViT-MAE Huge	0.411	0.386	**0.657**	0.508
ViT-MAE Base	DINOv2 Large	0.577	0.624	**0.732**	0.283
ViT-MAE Base	DINOv2 Giant	0.202	0.180	**0.580**	0.293
ViT-MAE Base	ViT-MAE Large	0.588	0.598	**0.850**	0.596
ViT-MAE Base	ViT-MAE Huge	0.149	0.417	**0.788**	0.571
ViT-MAE Huge	DINOv2 Giant	0.317	0.352	**0.614**	0.359

Table 3: **Vision model alignment performance** Reconstruction accuracy under HOT and pairwise OT, reported both in the standard (rotation-sensitive) and rotation-augmented (HOT+R) settings.

Results. Table 3 shows that, on its own, vanilla HOT does not consistently outperform pairwise OT in reconstruction accuracy. Consistent with this, the transport plans inferred by HOT (Figure 4 and Figures C.1 - C.10) only partially reveal layer-wise correspondences: in some cases, clear diagonal structure emerges, but in others the mappings are noisier. By contrast, the rotation-augmented variant (HOT+R) yields substantially higher reconstruction scores surpassing both vanilla HOT and pairwise vanilla and rotational baselines (Tables 3, C.1 and C.2). Importantly, the transport plans produced by HOT+R consistently exhibit strong hierarchical correspondences, recovering clean layer-to-layer alignment even in settings where vanilla HOT fails to do so. These findings indicate that incorporating rotation into the OT framework not only improves quantitative alignment quality but also produces more generalizable and interpretable mappings, particularly in domains like Vision Transformers where representations are known to be rotation-invariant.

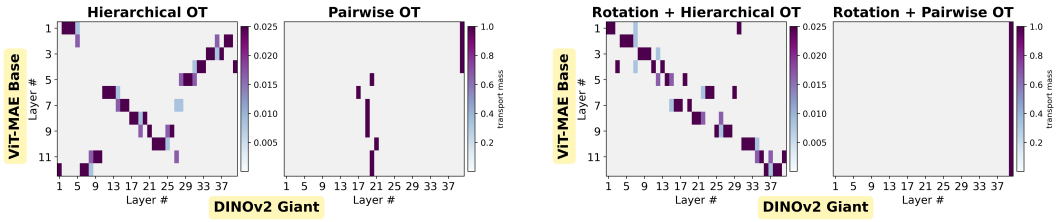

Figure 4: **Transport plans for vision model alignment.** ViT-MAE Base $\leftrightarrow$ DINOv2 Giant (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R). HOT+R captures geometric equivalences induced by rotations, yielding clearer correspondences than the rotation-sensitive variant.

4 Discussion

Hierarchical OT provides a principled framework for aligning two networks with arbitrary depths. By softly coupling all layer objectives, HOT (i) allows a neuron in one layer of a source network to align with a soft combination of units distributed across multiple layers of a target network, (ii) enforces global consistency that mitigates overfitting to noise in any single layer, and (iii) produces a single network-level—rather than layer-level—alignment score. In our experiments, HOT achieves higher alignment scores than greedy pairwise matches and yields intuitive transport plans that reveal hierarchical correspondences both in artificial networks and in the human brain.

Limitations. Despite these strengths, several limitations remain. First, HOT is computationally demanding: solving the inner OT scales as $O(n^3 \log n)$ for n neurons in a pair of layers, and the overall procedure is $O(L^2 n^3 \log n)$ for networks with L layers. This makes scaling to very wide or deep models challenging without further algorithmic improvements. Second, our evaluation was limited to a subset of models and brain datasets; follow-up analyses on more models and diverse neural data will be critical to assess generality. Finally, while HOT provides a descriptive alignment, it does not by itself explain *why* certain features converge across systems, or which computational principles drive these correspondences.

Future directions. Several extensions are natural. One avenue is to add a third hierarchical level to capture representational dynamics over the course of training, enabling alignment not only across neurons and layers but also across developmental or learning trajectories. Applications could include aligning models across different stages of training, or comparing biological learning to gradient descent. Another is to incorporate priors on the transport plan (e.g. smoothness) to guide alignment toward more interpretable solutions. Ultimately, as both biological and artificial networks grow in scale and complexity, methods like HOT that respect global structure while revealing network correspondences will be essential for understanding the universal principles governing intelligent systems.

REFERENCES

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. *Advances in neural information processing systems*, 34:225–236, 2021.

Zirui Chen and Michael F Bonner. Universal dimensions of visual representation. *Science Advances*, 11(27):eadw7697, 2025.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6(1):27755, 2016.

Jesús A De Loera and Edward D Kim. Combinatorics and geometry of transportation polytopes: An update. *Discrete geometry and algebraic combinatorics*, 625:37–76, 2013.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Robert Desimone and Stanley J Schein. Visual properties of neurons in area v4 of the macaque: sensitivity to stimulus form. *Journal of neurophysiology*, 57(3):835–868, 1987.

Robert Desimone, Thomas D Albright, Charles G Gross, and Charles Bruce. Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, 4(8):2051–2062, 1984.

Amil Dravid, Yossi Gandelsman, Alexei A Efros, and Assaf Shocher. Rosetta neurons: Mining the common units in a model zoo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1934–1943, 2023.

Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152: 184–194, 2017.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Šuppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*, 2025. doi: 10.48550/arXiv.2502.13595. URL https://arxiv.org/abs/2502.13595.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu,

Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng

Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Umut Güçlü and Marcel AJ Van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

Eghbal Hosseini, Colton Casto, Noga Zaslavsky, Colin Conwell, Mark Richardson, and Evelina Fedorenko. Universality of representation in biological and artificial neural networks. *bioRxiv*, 2024.

David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

Seyed-Mehdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLOS Computational Biology*, 2014. URL https://doi.org/10.1371/journal.pcbi.1003915.

Meenakshi Khosla and Alex H Williams. Soft matching distance: A metric on neural representations that captures single-neuron tuning. In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, pages 326–341. PMLR, 2024.

Meenakshi Khosla, Alex H Williams, Josh McDermott, and Nancy Kanwisher. Privileged representational axes in biological and artificial neural networks. *bioRxiv*, pages 2024–06, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

Philip May. Machine translated multilingual sts benchmark dataset. 2021. URL https://github.com/PhilipMay/stsb-multi-mt.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022.

Anitha Pasupathy and Charles E Connor. Shape representation in area v4: position-specific tuning for boundary conformation. *Journal of neurophysiology*, 86(5):2505–2519, 2001.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018.

Martin Schrimpf, Jonas Kubilius, Michael J Lee, N Apurva Ratan Murty, Robert Ajemian, and James J DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 2020. URL https://www.cell.com/neuron/fulltext/S0896-6273(20)30605-X.

Katherine R Storrs, Tim C Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus Kriegeskorte. Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of cognitive neuroscience*, 33(10):2044–2064, 2021.

John von Neumann. *A Certain Zero-sum Two-person Game Equivalent to the Optimal Assignment Problem*, pages 5–12. Princeton University Press, Princeton, 1953.

Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. *Advances in Neural Information Processing Systems*, 34:4738–4750, 2021.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

APPENDIX

A COMPARING LLM-LLM REPRESENTATIONS

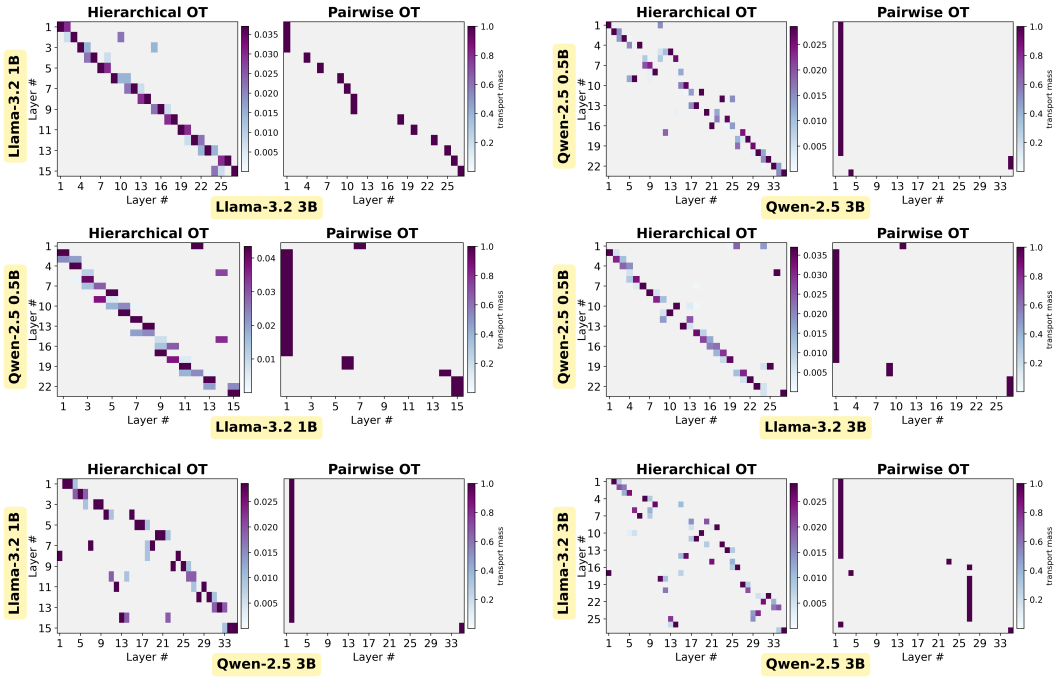

Figure A.1: **Transport plans across LLM families and scales.** Hierarchical OT (HOT) mappings are shown for six cross-model comparisons: (a) LLaMA-3.2 1B ↔ LLaMA-3.2 3B, (b) Qwen-2.5 0.5B ↔ Qwen-2.5 3B, (c) Qwen-2.5 0.5B ↔ LLaMA-3.2 1B, (d) Qwen-2.5 0.5B ↔ LLaMA-3.2 3B, (e) LLaMA-3.2 1B ↔ Qwen-2.5 3B, and (f) LLaMA-3.2 3B ↔ Qwen-2.5 3B. HOT uncovers structured, near-diagonal correspondences that persist across both intra-family (a,b) and cross-family (c-f) alignments, illustrating its robustness compared to pairwise OT.

B Representation Similarity between Vision Cortex Response

Model 1	Model 2	HOT Metric	Random (Perm-P)	Single-Best OT	Pairwise Best OT
Subject A	Subject B	0.306 ± 0.007	0.258 ± 0.008	0.306 ± 0.007	0.308 ± 0.007
Subject A	Subject C	0.227 ± 0.008	0.184 ± 0.008	0.227 ± 0.008	0.232 ± 0.009
Subject A	Subject D	0.248 ± 0.014	0.212 ± 0.016	0.248 ± 0.014	0.253 ± 0.014
Subject B	Subject C	0.237 ± 0.008	0.199 ± 0.006	0.237 ± 0.008	0.237 ± 0.009
Subject C	Subject D	0.235 ± 0.010	0.196 ± 0.009	0.235 ± 0.010	0.239 ± 0.008
Subject B	Subject D	0.248 ± 0.013	0.217 ± 0.011	0.248 ± 0.013	0.254 ± 0.012

Table B.1: **Visual cortex alignment performance.** Comparison of HOT against baseline methods, evaluated by reconstruction correlation on held-out fMRI responses: mean $\pm$ standard deviation across seeds.

B.1 Transport plans for OT based mapping pairs

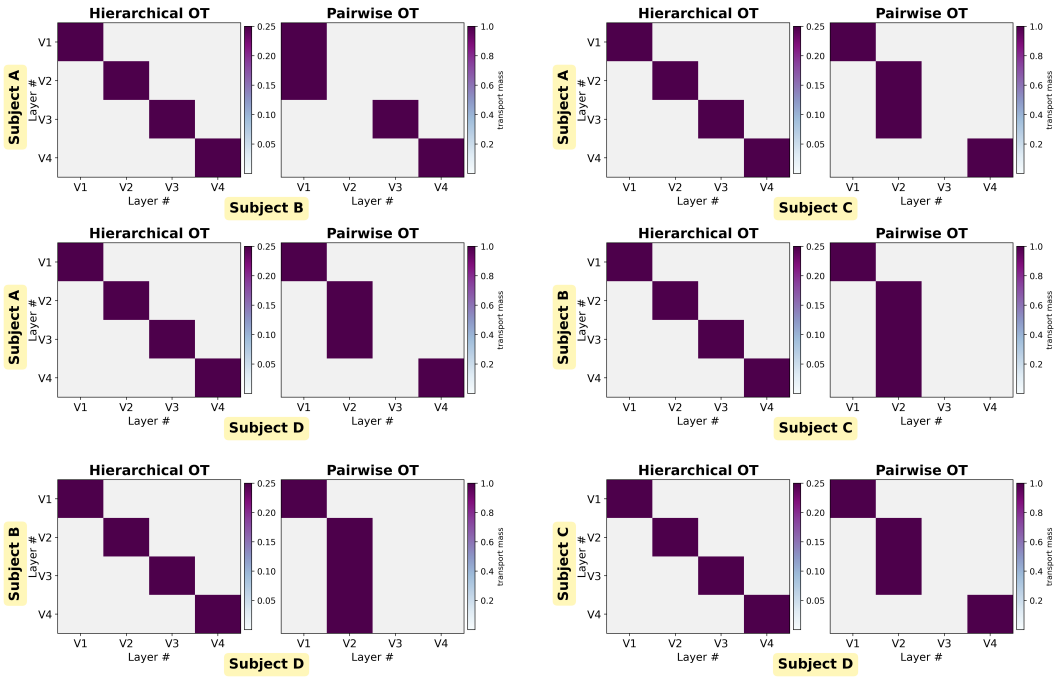

Figure B.1: **Transport plans for OT-based mapping pairs.** Pairwise OT mappings across all subject pairs: (a) Subject A $\leftrightarrow$ Subject B, (b) Subject A $\leftrightarrow$ Subject C, (c) Subject A $\leftrightarrow$ Subject D, (d) Subject B $\leftrightarrow$ Subject C, (e) Subject B $\leftrightarrow$ Subject D, and (f) Subject C $\leftrightarrow$ Subject D. HOT recovers structured region-to-region correspondences that are absent in pairwise OT.

B.2 Transport plans for Linear Predictivity based mapping pairs

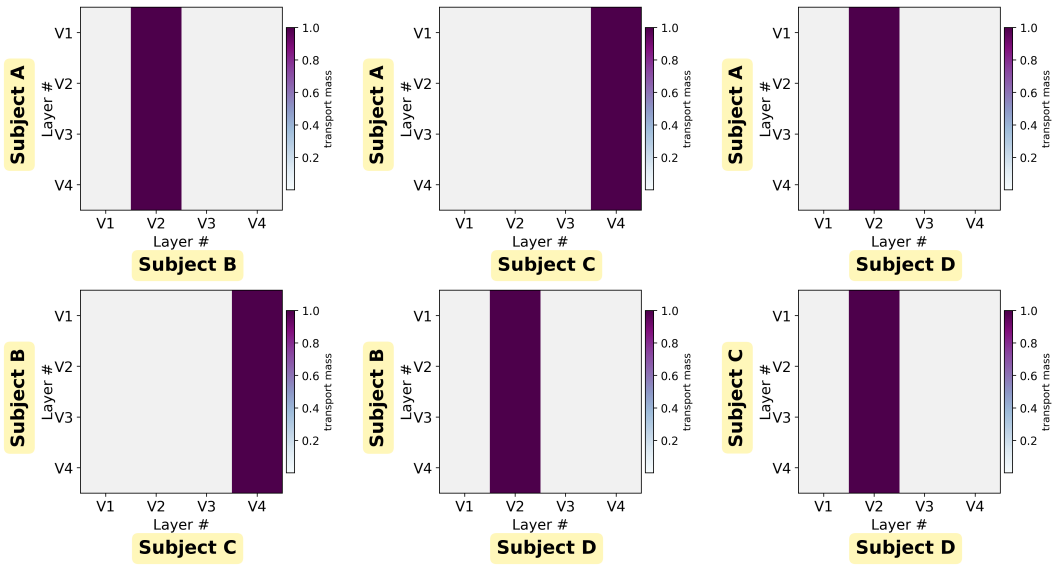

Figure B.2: **Transport plans for linear predictivity-based mapping pairs.** Pairwise OT mappings learned under linear predictivity constraints for all subject pairs: (a) Subject A ↔ Subject B, (b) Subject A ↔ Subject C, (c) Subject A ↔ Subject D, (d) Subject B ↔ Subject C, (e) Subject B ↔ Subject D, and (f) Subject C ↔ Subject D. Unlike hierarchical OT, linear predictivity-based pairwise OT fails to recover structured layer-wise correspondences, highlighting the importance of hierarchical optimization for learning robust and generalizable mappings across subjects.

C Representation Similarity between Vision Models

Model 1	Model 2	HOT Metric	Random (Perm-P)	Single-Best OT	Pairwise Best OT
DINOv2 Small	ViT-MAE Base	0.289	0.259	0.289	**0.301**
DINOv2 Small	DINOv2 Large	**0.353**	0.298	0.312	0.340
DINOv2 Small	DINOv2 Giant	**0.466**	0.385	0.413	0.433
DINOv2 Small	ViT-MAE Large	**0.381**	0.348	0.350	0.354
DINOv2 Small	ViT-MAE Huge	**0.411**	0.372	0.359	0.386
ViT-MAE Base	DINOv2 Large	0.577	0.394	0.531	**0.624**
ViT-MAE Base	DINOv2 Giant	**0.202**	0.197	0.148	0.180
ViT-MAE Base	ViT-MAE Large	0.588	0.528	0.539	**0.598**
ViT-MAE Base	ViT-MAE Huge	0.149	0.116	0.181	**0.417**
ViT-MAE Huge	DINOv2 Giant	0.317	0.261	0.293	**0.352**

Table C.1: HOT Metric with its baseline comparisons.

Model 1	Model 2	HOT + R	Single-Best + R	Pairwise Best + R
DINOv2 Small	ViT-MAE Base	**0.600**	**0.600**	0.526
DINOv2 Small	DINOv2 Large	**0.778**	0.708	0.394
DINOv2 Small	DINOv2 Giant	**0.790**	0.746	0.418
DINOv2 Small	ViT-MAE Large	**0.633**	0.599	0.509
DINOv2 Small	ViT-MAE Huge	**0.657**	0.614	0.508
ViT-MAE Base	DINOv2 Large	**0.732**	0.712	0.283
ViT-MAE Base	DINOv2 Giant	**0.580**	0.605	0.293
ViT-MAE Base	ViT-MAE Large	**0.850**	0.848	0.596
ViT-MAE Base	ViT-MAE Huge	**0.788**	0.760	0.571
ViT-MAE Huge	DINOv2 Giant	**0.614**	0.582	0.359

Table C.2: Rotation (HOT + R) with its baseline comparisons.

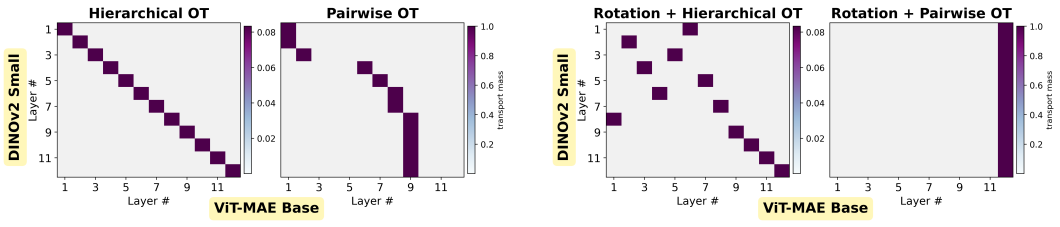

Figure C.1: **Transport plans for vision model alignment.** DINOv2 Small ↔ ViT-MAE Base (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

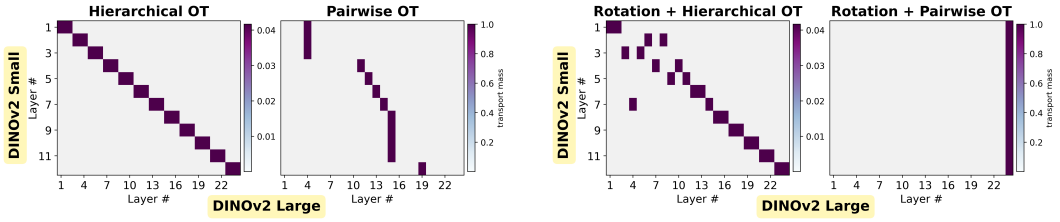

Figure C.2: **Transport plans for vision model alignment.** DINOv2 Small ↔ DINOv2 Large (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

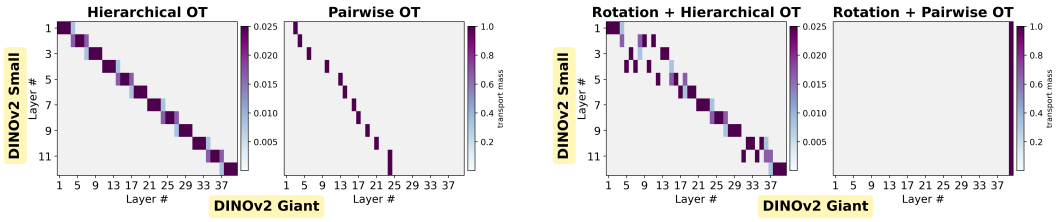

Figure C.3: **Transport plans for vision model alignment.** DINOv2 Small ↔ DINOv2 Giant (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

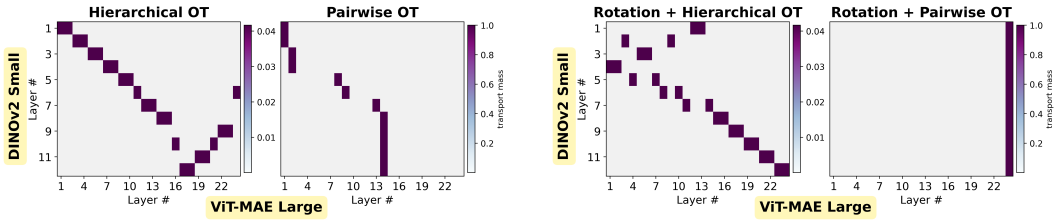

Figure C.4: **Transport plans for vision model alignment.** DINOv2 Small ↔ ViT-MAE Large (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

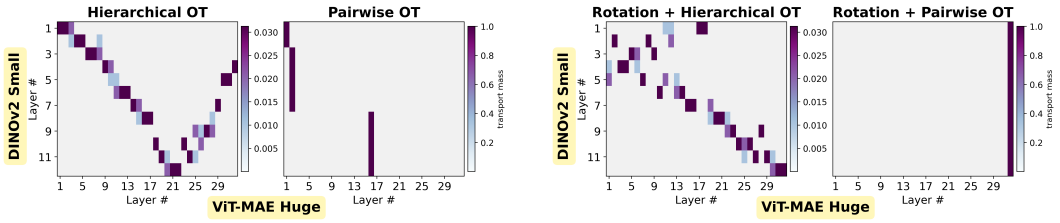

Figure C.5: **Transport plans for vision model alignment.** DINOv2 Small ↔ ViT-MAE Huge (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

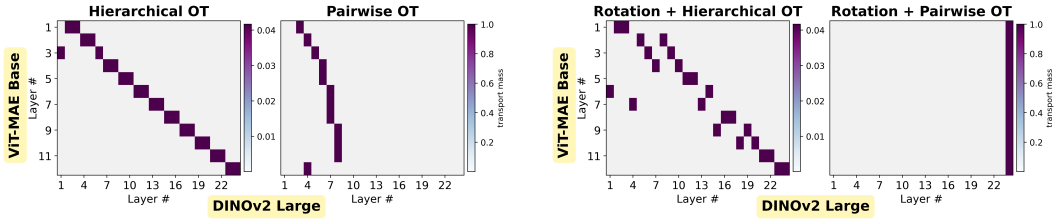

Figure C.6: **Transport plans for vision model alignment.** ViT-MAE Base ↔ DINOv2 Large (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

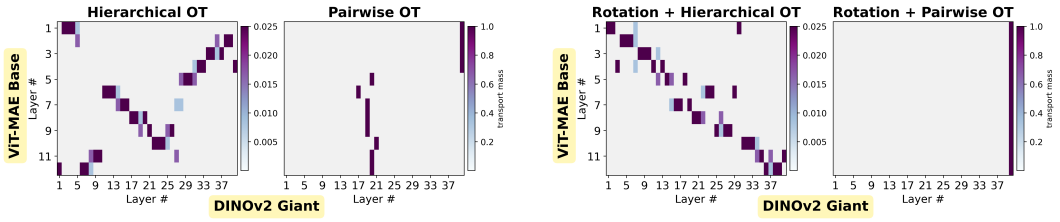

Figure C.7: **Transport plans for vision model alignment.** ViT-MAE Base ↔ DINOv2 Giant (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

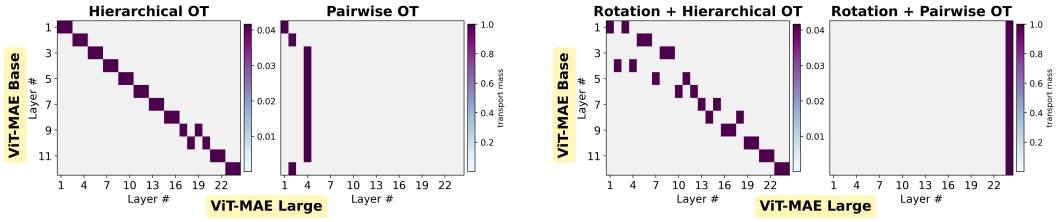

Figure C.8: **Transport plans for vision model alignment.** ViT-MAE Base ↔ ViT-MAE Large (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

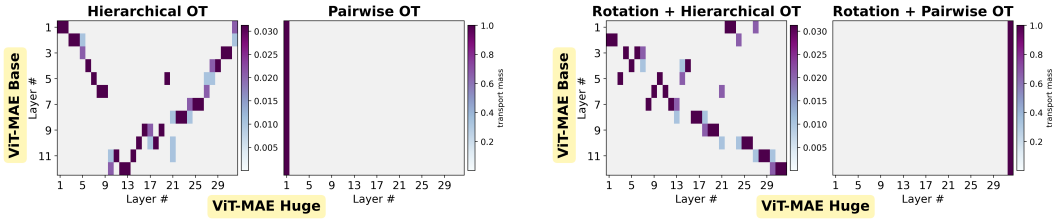

Figure C.9: **Transport plans for vision model alignment.** ViT-MAE Base ↔ ViT-MAE Huge (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

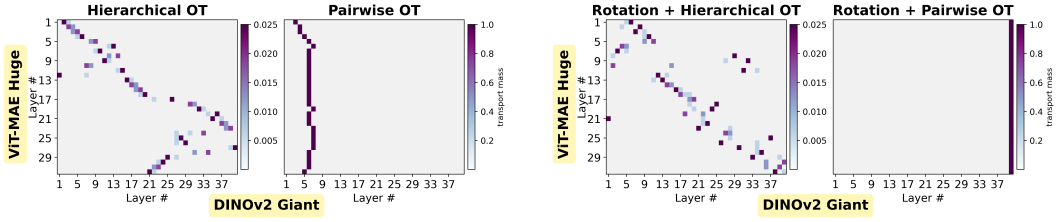

Figure C.10: **Transport plans for vision model alignment.** ViT-MAE Huge ↔ DINOv2 Giant (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

D Representation Similarity between Vision Models and Vision Cortex

Experimental Setup. We assess representational similarity between human visual cortex and vision transformers by comparing fMRI responses with model activations elicited by the same set of stimuli. Specifically, we use responses from the Natural Scenes Dataset ((Allen et al., 2022)), focusing on 1,000 shared images viewed by participants in the fMRI experiment. The same images are presented to pretrained vision transformers, and layer-wise representations are extracted by averaging patch embeddings across each input. Following the approach used in the cortex–cortex analysis, we treat distinct visual areas (V1–V4) as "layers" and individual voxels as "neurons." We then compute both HOT and its rotation-augmented variant (HOT+R) to align cortical responses with model representations. This design enables a direct comparison of hierarchical organization across biological and artificial systems under matched visual input.

Model 1	Model 2	HOT Metric	Pairwise Best OT	HOT + R	Pairwise Best + R
Subject	DINOv2 Base	0.092	0.084	**0.145**	0.094
Subject	DINOv2 Giant	0.090	0.071	**0.163**	0.110
Subject	ViT-MAE Base	0.127	0.114	**0.151**	0.121
Subject	ViT-MAE Huge	0.072	0.099	**0.154**	0.122

Table D.1: Results on HOT metric vs. baselines (test split).

Results. Table D.1 reports reconstruction scores for HOT, HOT+R, and corresponding baselines averaged across the four subjects. Among all methods, HOT+R achieves the highest reconstruction score, indicating that incorporating rotation is critical for capturing shared structure between cortical and model representations. Subject-specific results (Tables D.2 and D.3) further confirm this trend, with HOT+R consistently outperforming both vanilla HOT and pairwise baselines. Transport plans visualized in Figures D.1– D.16 reveal partial but inconsistent layer-wise correspondences: in some cases, early cortical regions align with early model layers and higher regions map to deeper layers, while in others the mappings appear noisier.

Model 1	Model 2	HOT Metric	Random (Perm-P)	Single-Best OT	Pairwise Best OT
Subject A	DINOv2 Giant	0.085	0.087	0.046	0.063
Subject A	DINOv2 Base	0.081	0.072	0.075	0.080
Subject A	ViT-MAE Huge	0.065	0.057	0.063	0.105
Subject A	ViT-MAE Base	0.141	0.131	0.123	0.124
Subject B	DINOv2 Giant	0.087	0.089	0.054	0.065
Subject B	DINOv2 Base	0.088	0.080	0.075	0.082
Subject B	ViT-MAE Huge	0.077	0.072	0.073	0.093
Subject B	ViT-MAE Base	0.115	0.109	0.101	0.104
Subject C	DINOv2 Giant	0.115	0.124	0.072	0.092
Subject C	DINOv2 Base	0.115	0.108	0.095	0.100
Subject C	ViT-MAE Huge	0.092	0.088	0.082	0.114
Subject C	ViT-MAE Base	0.137	0.134	0.121	0.124
Subject D	DINOv2 Giant	0.073	0.082	0.049	0.063
Subject D	DINOv2 Base	0.083	0.078	0.072	0.072
Subject D	ViT-MAE Huge	0.054	0.049	0.046	0.086
Subject D	ViT-MAE Base	0.114	0.111	0.101	0.106

Table D.2: Results on HOT metric vs. baselines (test split).

Model 1	Model 2	HOT + R	Single-Best + R	Pairwise Best + R
Subject A	DINOv2 Giant	0.161	0.140	0.099
Subject A	DINOv2 Base	0.146	0.140	0.083
Subject A	ViT-MAE Huge	0.165	0.151	0.127
Subject A	ViT-MAE Base	0.157	0.151	0.124
Subject B	DINOv2 Giant	0.147	0.132	0.096
Subject B	DINOv2 Base	0.127	0.120	0.084
Subject B	ViT-MAE Huge	0.125	0.109	0.104
Subject B	ViT-MAE Base	0.128	0.118	0.101
Subject C	DINOv2 Giant	0.205	0.183	0.147
Subject C	DINOv2 Base	0.185	0.181	0.128
Subject C	ViT-MAE Huge	0.189	0.179	0.151
Subject C	ViT-MAE Base	0.186	0.180	0.152
Subject D	DINOv2 Giant	0.139	0.123	0.099
Subject D	DINOv2 Base	0.121	0.114	0.081
Subject D	ViT-MAE Huge	0.135	0.122	0.107
Subject D	ViT-MAE Base	0.134	0.129	0.109

Table D.3: Results on HOT metric vs. baselines (test split).

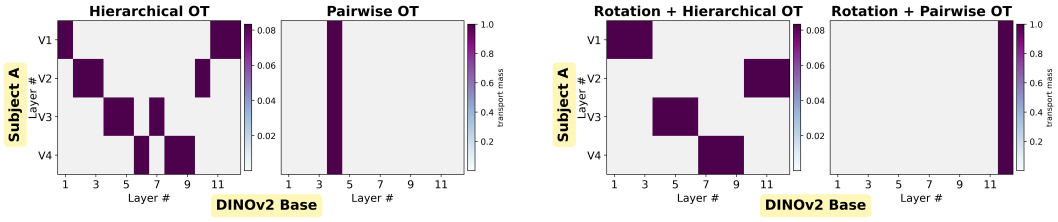

Figure D.1: Subject A ↔ DINOv2 Base (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

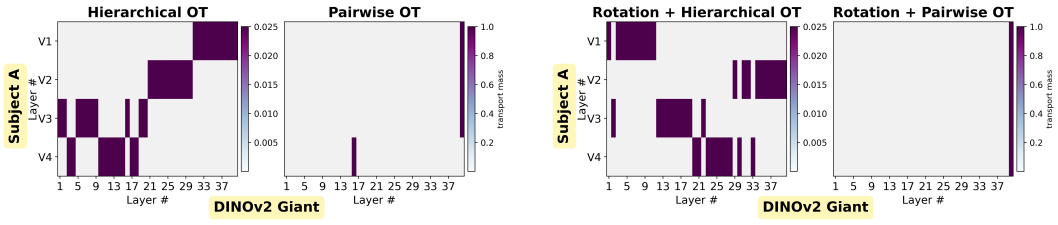

Figure D.2: Subject A ↔ DINOv2 Giant (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

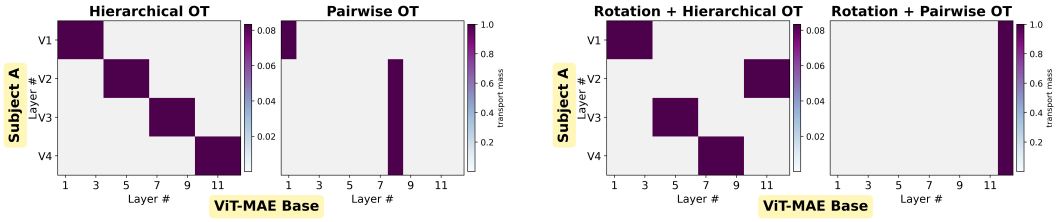

Figure D.3: Subject A ↔ ViT-MAE Base (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

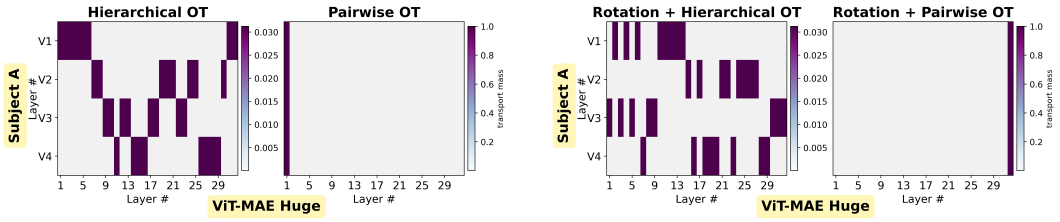

Figure D.4: Subject A ↔ ViT-MAE Huge (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

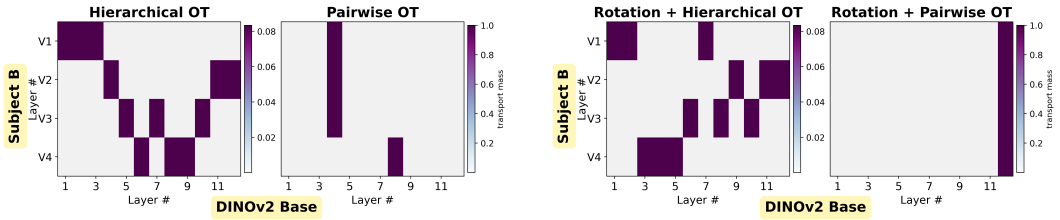

Figure D.5: Subject B ↔ DINOv2 Base (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

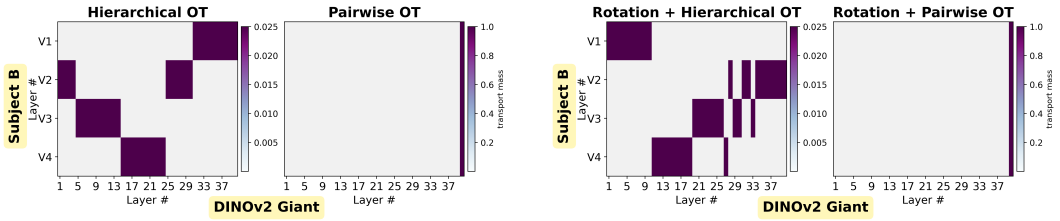

Figure D.6: Subject B ↔ DINOv2 Giant (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

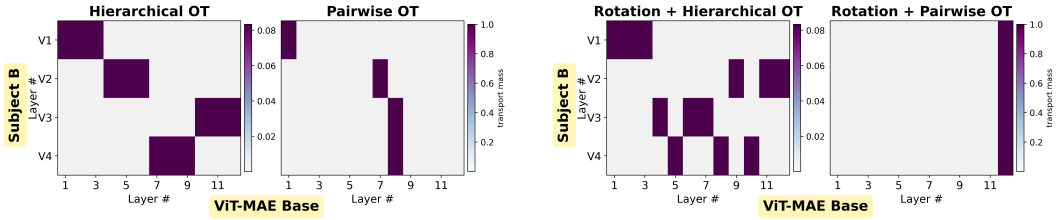

Figure D.7: Subject B ↔ ViT-MAE Base (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

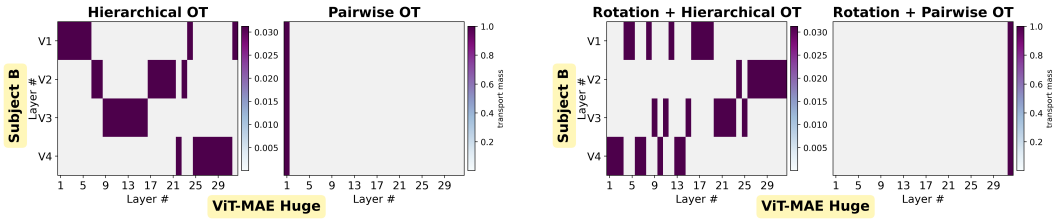

Figure D.8: Subject B ↔ ViT-MAE Huge (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

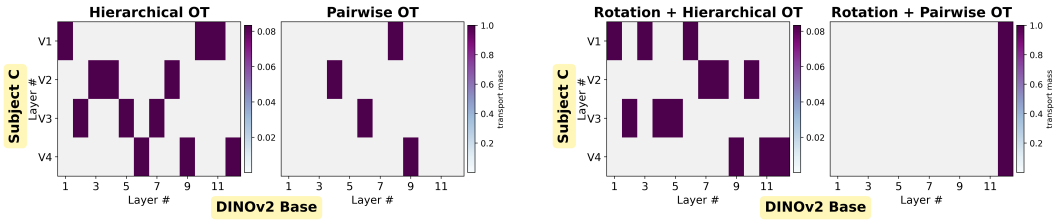

Figure D.9: Subject C ↔ DINOv2 Base (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

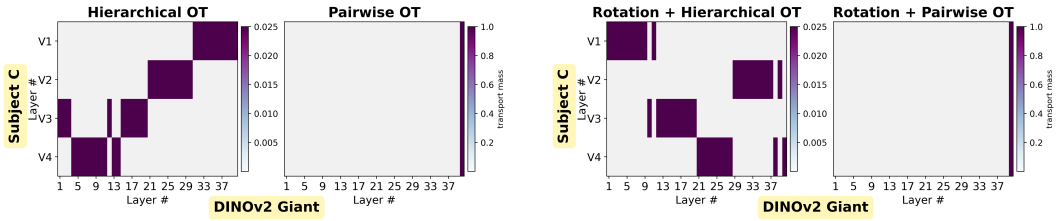

Figure D.10: Subject C ↔ DINOv2 Giant (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

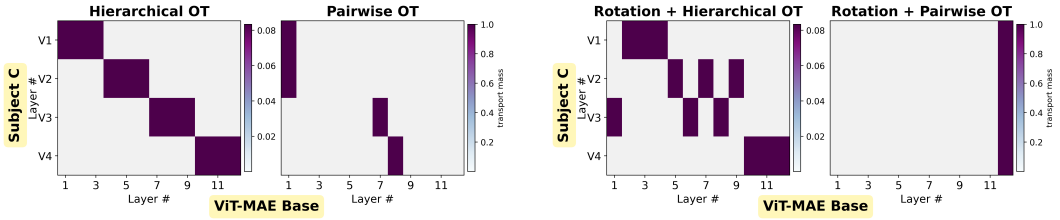

Figure D.11: Subject C ↔ ViT-MAE Base (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

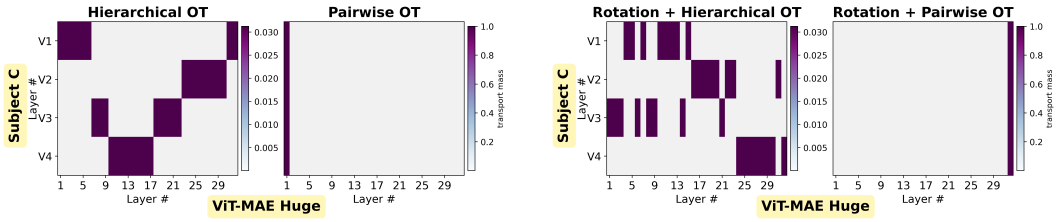

Figure D.12: Subject C ↔ ViT-MAE Huge (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

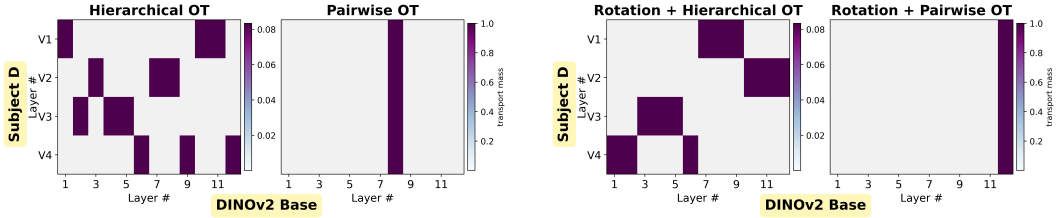

Figure D.13: Subject D ↔ DINOv2 Base (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

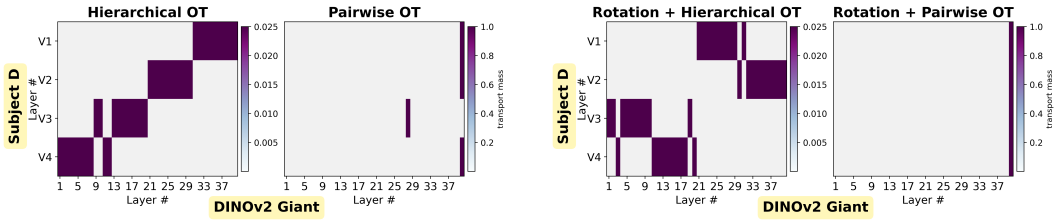

Figure D.14: Subject D ↔ DINOv2 Giant (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

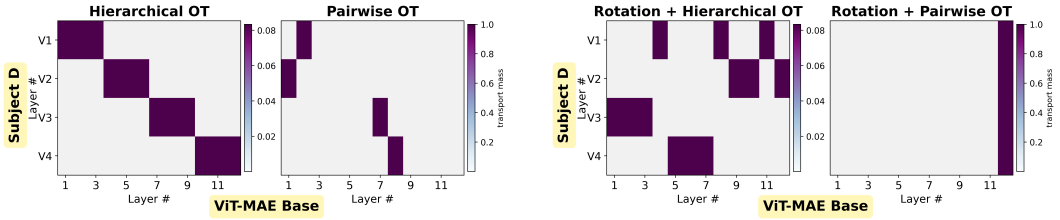

Figure D.15: Subject D ↔ ViT-MAE Base (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

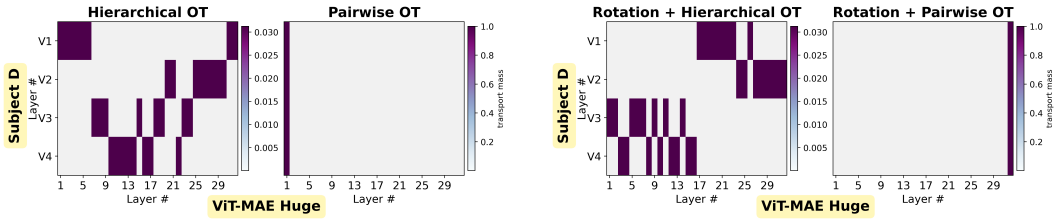

Figure D.16: Subject D ↔ ViT-MAE Huge (a) without rotation (HOT) and (b) with rotation augmentation (HOT+R).

E USE OF LARGE LANGUAGE MODELS

LLMs were used in this work to assist with writing tasks, specifically for ensuring grammatical correctness and enhancing the clarity of the text.