WIRED TO REASON: CONNECTOME RESERVOIRS ALIGN WITH VISION-LANGUAGE MODELS ON RAVEN'S MATRICES.

Anonymous authors
Paper under double-blind review

ABSTRACT

The human brain and large neural networks have independently converged on surprisingly similar internal representations, a finding typically attributed to learning at scale. But what if a part of this convergence requires no gradient learning in the recurrent operators? We test this using a *connectome-wired reservoir*: a recurrent dynamical system whose bilinear operators are derived entirely from empirical brain connectivity matrices, with *no learning in the recurrent weights*. Remarkably, the internal states this system builds during visual reasoning align with the layer-wise representations of a trained vision-language model in a structured and interpretable way: alignment varies systematically across recurrent states and VLM depth, with stronger alignment for correct trials in several model–metric combinations, without gradient-based optimization. This alignment is not a statistical artifact: empirical FC exhibits modest but consistently stronger alignment than the matched random control, while temporal-memory removal alters the alignment structure. Moreover, reservoirs instantiated with connectomes from both neurotypical and autistic individuals produce consistent visual reasoning performance that substantially exceeds random-weight baselines, indicating that this advantage generalizes across both cohorts rather than being tied to any particular population or diagnostic profile; structure-matched null controls further indicate that the empirical connectome's regional arrangement, not any single isolated property, underlies the alignment advantage. Taken together, our results suggest that the partial convergence between biological and vision-language models (VLMs) is not exclusively learned: FC-derived recurrent organization provides a structural inductive bias for partial CWR–VLM representational correspondence. Here, connectome-wired refers specifically to the use of an empirical brain connectivity matrix to determine the fixed recurrent dynamics of the reservoir. No measured neural activity is used as a representation target: all alignment analyses compare reservoir states with representations from trained vision-language models.

1 INTRODUCTION

The open question. A striking finding in computational neuroscience is that large neural networks, despite being trained without any biological objective, develop internal representations that partially resemble those of the human brain (Hosseini et al., 2024; Caucheteux et al., 2022; Shen et al., 2025). For vision, layer-wise activations of deep convolutional networks align with the hierarchy of responses in the ventral visual stream (Eickenberg et al., 2017). For language, a sharper correspondence has been documented: the depth of a language model layer predicts the *latency* at which its activations best match cortical responses recorded with magnetoencephalography (MEG), revealing a shared *sequence of computations* (Raugel et al., 2025). The standard interpretation is that this convergence follows from learning at scale: models trained on rich data converge on representations that happen to resemble those the brain has shaped through evolution and experience (Huh et al., 2024). This interpretation conflates two distinct questions. Is gradient-based learning *necessary* for this convergence? Or is some portion of it already encoded in the structural organization of biological connectivity, *prior to any training*? These questions are empirically indistinguishable using trained models alone; this paper provides a way to separate them.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This distinction matters for how the rest of the paper should be read: prior work of this kind compares trained artificial networks to *measured* brain activity (fMRI, MEG) recorded while a person performs a task. Our comparison is different in kind. We use an empirical brain connectivity matrix only to fix the recurrent dynamics of an otherwise untrained reservoir, and we compare the internal states that this reservoir produces to the layer-wise representations of a trained vision-language model. No neural recordings enter the analysis at any point. Throughout the paper, "brain-derived" and "connectome-wired" refer specifically to the origin of the reservoir's fixed connectivity.

Our approach: connectome-wired reservoir as a probe. Reservoir computing (Nakajima & Fischer, 2021) provides a controlled setting in which the role of network structure can be isolated. In echo-state networks (ESN) (Jaeger, 2001) and liquid-state machines (LSM) (Maass et al., 2002), the recurrent weights are fixed at initialization and never updated, so the representations that emerge are determined by network topology rather than learned parameters. We leverage this by instantiating the reservoir with empirical brain connectivity matrices derived from resting-state fMRI, replacing standard random initialization (Lukoševičius & Jaeger, 2009), yielding a dynamical system whose computation is entirely driven by biological structure. To test whether such a system supports structured relational computation, and whether the representations it builds bear a systematic resemblance to those of trained artificial systems, we need a task that isolates abstract reasoning from perceptual pattern matching. We use Raven's Progressive Matrices (RPM) (Zhang et al., 2019), where each problem requires inferring a latent relational rule from context panels and applying it to select the correct candidate. This lets us track how internal states evolve as the system processes relational information and compare these states directly to the layer-wise representations of a vision-language model solving the same task. We call this system *connectome-wired reservoir (CWR)*: the brain connectivity matrix directly defines the recurrent dynamics, while only the readout is trained, which lets us attribute the resulting representations to connectivity structure rather than learned weights. We verify this separation directly: because the recurrent operators never receive a gradient, the reservoir states we analyze are identical by construction whether the readout is untrained, trained, or trained with a different objective (Appendix B). We design our experiments around three hypotheses:

> ***Hypothesis 1.*** *Alignment between connectome-wired reservoir states and VLM layers varies systematically with reasoning step and model depth.*
>
> ***Hypothesis 2.*** *Alignment between reservoir states and VLM layers differs between correctly and incorrectly solved visual reasoning problems.*
>
> ***Hypothesis 3.*** *Reservoirs with brain-derived connectivity behave differently from randomly connected reservoirs of the same size.*

2 Related Work

Reservoir computing and brain connectivity. Echo state networks (ESN) (Jaeger, 2001) and liquid state machines (LSM) (Maass et al., 2002) showed that a fixed, random recurrent network is sufficient for a linear readout to solve complex temporal tasks, shifting the computational burden from learned dynamics to fixed structure. Population-level connectivity-template construction has also been explored by aggregating subject-level networks across cohorts (Soussia et al., 2025). Suárez et al. (2021) introduced human structural connectomes as neuromorphic reservoirs and showed that network topology, specifically the modular organization of large-scale functional systems, drives memory capacity beyond matched random networks at the critical dynamical regime; Suárez et al. (2024) extended this into a general-purpose toolbox, and empirical connectomes can influence reservoir dynamics, while some conventional ESNs achieve comparable performance (Damicelli et al., 2022). Across this work, two constraints have remained constant: reservoir states are vector-valued, discarding the two- and three-dimensional spatial structure of the input, and evaluation has been restricted to memory and signal-processing benchmarks, never to abstract relational reasoning. We remove both constraints simultaneously: by wiring the recurrent operator with empirical brain connectivity, we evaluate connectome-based reservoirs on structured relational reasoning and ask whether the representations they build in doing so geometrically resemble those of a trained vision-language model, a question that, to our knowledge, has not previously been posed.

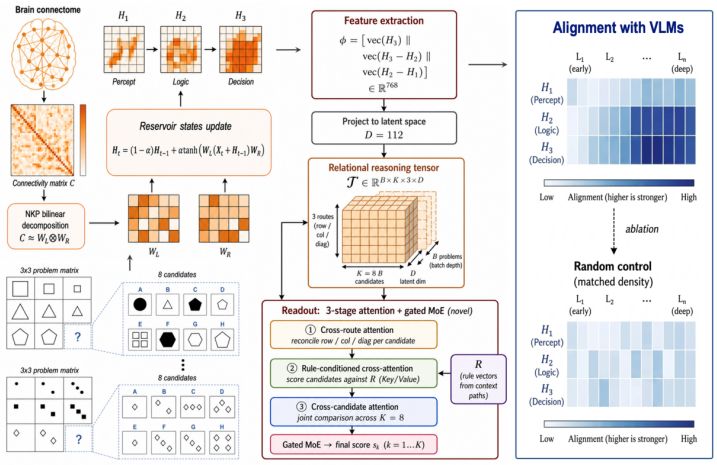


Figure 1: *Overview of the proposed CWR framework for visual reasoning*. A brain connectome is factorized into reservoir operators that recursively integrate candidate representations into hierarchical states (H_1–H_3: early, intermediate, and terminal integration). These states are projected into a relational reasoning tensor for candidate scoring and layer-wise alignment with VLM representations.

Representational alignment between trained models and the brain. The goal-driven modeling framework (Yamins & DiCarlo, 2016) posits that networks optimized for naturalistic tasks develop representations resembling those in sensory cortex, confirmed across visual (Yamins et al., 2014; Eickenberg et al., 2017), auditory (Kell et al., 2018), and language (Schrimpf et al., 2021; Caucheteux et al., 2022) systems. Raugel et al. (2025) showed that LLM layer depth predicts the latency of peak MEG alignment during speech listening, revealing a shared *ordering* of computations across transformer and recurrent architectures. Notably, untrained GPT2-xl achieves approximately 51% average predictivity, around 20 percentage points below its trained counterpart (Schrimpf et al., 2021), showing that some brain-predictive structure can be present before training. This raises the question our work addresses: does alignment persist, and remain interpretable, when the topology is not a randomly initialized design but is instead fixed by the functional wiring of the human brain? This literature compares trained networks to *measured* brain activity, whereas we compare a connectome-derived, untrained reservoir to a trained VLM; we build on its central logic (architecture alone can carry representational structure) without repeating its brain-recording methodology.

Abstract visual reasoning and VLM alignment. The RAVEN benchmark (Zhang et al., 2019) is a standard testbed for relational visual reasoning; models solving it can exploit statistical biases in the answer set rather than relational structure (Hu et al., 2021). Representational alignment between reasoning systems has so far been studied through the lens of trained-model comparison (Webb et al., 2023). We instead use RAVEN to compare a fixed, gradient-free connectome-derived system to trained VLMs, contrasting learned representations with structure-driven dynamics.

3 Method

We study whether the internal reasoning states of a reservoir whose recurrent weights are fixed by a brain connectivity matrix align with the layer-wise representations of a trained VLM during abstract visual reasoning (Figure 1). Our framework *connectome-wired reservoir* (CWR) is composed of a biologically-wired dynamical system whose recurrent operators receive no gradient, and followed by a representational alignment protocol that compares its states to VLM representations across depth.

3.1 Problem setup and spatial decomposition

We use Raven's Progressive Matrices (RPM) (Zhang et al., 2019) as the reasoning task. Each RAVEN problem presents a 3×3 grid governed by a hidden relational rule; one panel is missing and must be selected from $K = 8$ candidates (Figure 1). Solving RPM requires identifying latent rules from

context panels rather than matching surface statistics. Each visual panel is z-scored, downsampled via bicubic interpolation, then compressed to a compact matrix-valued input via fixed bilinear projections. We decompose each problem into **29 structured reasoning paths**: *Context Paths* (5 paths spanning the first two rows, first two columns, and anti-diagonal) and *Candidate Paths* (24 paths, $K = 8$ candidates $\times$ 3 spatial directions), decomposing RPM into parallel independent reasoning streams (Appendix C).

3.2 Connectome-Wired Reservoir (CWR)

From connectome to bilinear operator. Let $\mathbf{C} \in \mathbb{R}^{R\times R}$ be a brain connectivity matrix between R brain regions. We embed $\mathbf{C}$ as the recurrent operator via Bilinear Nearest Kronecker Product (NKP) decomposition (Van Loan & Pitsianis, 1993):

$$\mathbf{C} \approx \mathbf{W}_L \otimes \mathbf{W}_R, \qquad \mathbf{W}_L, \mathbf{W}_R \in \mathbb{R}^{d\times d}, \tag{1}$$

where $(\mathbf{W}_L, \mathbf{W}_R)$ minimizes $\|\mathbf{C} - \mathbf{W}_L \otimes \mathbf{W}_R\|_F$ in closed form. This simultaneously adapts $\mathbf{C}$ from parcellation resolution R to state dimension d, and yields separable bilinear operators that act along both spatial axes of each visual panel, directly encoding the topological structure of $\mathbf{C}$ in the recurrence. Both operators are fixed throughout.

Reservoir dynamics. Each $n\times n$ visual panel is compressed to a $d\times d$ matrix-valued input via a fixed bilinear projection $\mathbf{X}_{\text{proj}} = \mathbf{P}_L\mathbf{X}_{\text{img}}\mathbf{P}_R$, preserving the panel's 2D spatial layout. For each of the 29 paths, the reservoir state $\mathbf{H}_t \in \mathbb{R}^{d\times d}$ evolves over the sequential steps ($t = 1, 2, 3$) as:

$$\mathbf{H}_t = (1-\alpha)\mathbf{H}_{t-1} + \alpha\ \tanh\big(\mathbf{W}_L(\mathbf{X}_t + \mathbf{H}_{t-1})\mathbf{W}_R\big), \tag{2}$$

where $\alpha \in (0, 1]$ is the leak rate. $\mathbf{W}_L$ and $\mathbf{W}_R$ are fixed and connectome-derived. At $\alpha = 1$, leaky integration is eliminated, removing the explicit blending of past states. This regime, which we term *no-leak*, eliminates the temporal integration mechanism and provides the natural ablation point in (§4.1). The sequential states $\{\mathbf{H}_1, \mathbf{H}_2, \mathbf{H}_3\}$ carry distinct computational roles: early integration, intermediate integration, and terminal integration.

3.3 Relational readout and training

High-order feature extraction. Rather than using a single terminal state, the full temporal trajectory of each path is captured by concatenating three complementary representations:

$$\boldsymbol{\phi} = \big[\mathrm{vec}(\mathbf{H}_3) \parallel \mathrm{vec}(\mathbf{H}_3 - \mathbf{H}_2) \parallel \mathrm{vec}(\mathbf{H}_2 - \mathbf{H}_1)\big], \tag{3}$$

encoding the final state, late-stage logical convergence, and early-stage logical initiation respectively. This vector is projected into a shared latent space D via a learned linear operator; *only this projection and all subsequent components receive gradient updates*, the reservoir itself is fixed.

Differential relational modeling. In latent space, the framework extracts a *rule vector* $\mathbf{r}_r = \mathbf{h}_{r,2} - \mathbf{h}_{r,1}$ encoding the contextual relational transformation along route r, and a *candidate vector* $\mathbf{c}_r^{(k)} = \mathbf{h}_{r,3}^{(k)} - \mathbf{h}_{r,2}$ encoding the transformation induced by candidate k. Their concatenation is passed through a learned projection, and rule vectors are preserved in a standalone *logical benchmark tensor* $\mathcal{R}$ used to condition the subsequent attention stage. For the diagonal route, a dedicated *step-wise state analogy* decomposes reasoning into a corner-to-center step (anti-diagonal rule) and a center-to-answer step (main-diagonal candidate), handling the geometric fact that the center panel simultaneously anchors both diagonal axes. The three route-wise feature blocks are stacked into a single *relational reasoning tensor* $\mathcal{T}_{\text{rel}} \in \mathbb{R}^{B\times K\times 3\times D}$ (full construction in Appendix B.2).

Hierarchical attention and decision. $\mathcal{T}_{\text{rel}}$ is processed by three progressive attention stages: (1) cross-route self-attention reconciles the row, column, and diagonal perspectives for each candidate; (2) rule-conditioned cross-attention uses $\mathcal{R}$ as key/value to score each candidate against the inferred rule; (3) cross-candidate self-attention enables joint comparison across all K options. The output is augmented with explicit geometric scalars (cosine similarity, norm ratio, scaling error) per route before a *gated mixture-of-experts* head produces the final score:

$$s_k^{\text{final}} = \sum_{r\in\{\text{row,col,diag}\}} s_{k,r}\cdot w_r, \tag{4}$$

where $s_{k,r} \in \mathbb{R}$ denote the scalar score for candidate k along route r and and $w_r \in [0,1]$ with $\sum_r w_r = 1$ are adaptive gate weights generated by a small network conditioned on the global feature distribution. The readout is trained by minimizing:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_r \mathcal{L}_{\text{rank}} + \frac{\lambda_c}{P(P-1)} \sum_{p \neq q} D_{\text{Jeff}}(\pi^{(p)} \| \pi^{(q)}), \tag{5}$$

where $\mathcal{L}_{\text{CE}}$ is cross-entropy, $\mathcal{L}_{\text{rank}}$ is a margin-based ranking loss, and the third term enforces cross-route consistency via the Jeffreys divergence $D_{\text{Jeff}}(p\|q) = \frac{1}{2}[D_{\text{KL}}(p\|q) + D_{\text{KL}}(q\|p)]$ over all $P(P-1) = 6$ ordered route pairs, reflecting the structural assumption that a single relational rule governs all three spatial directions (Zhang et al., 2019). More details are in Appendix D.

3.4 Representational alignment protocol

The brain reservoir and VLM solve the same RAVEN problems under identical candidate ordering across N matched instances per task. For each VLM layer ℓ, global mean pooling over token positions yields a representation vector $\mathbf{s}_\ell \in \mathbb{R}^{d_{\text{VLM}}}$ per problem. From the reservoir, each reasoning stage $t \in \{1,2,3\}$ and route r contributes the vectorized state $\mathbf{h}_t^{(r)} = \text{vec}(\mathbf{H}_{t,r}) \in \mathbb{R}^{d^2}$. Stacking over N instances produces a $3 \times L$ alignment grid of (reservoir stage, VLM layer) scores per route, where L is the number of transformer layers. We quantify alignment using two measures:

Centered Kernel Alignment (CKA) (Kornblith et al., 2019) measures global covariance alignment between the two representation spaces. Given the $N \times d^2$ reservoir activation matrix $\mathbf{X}$ and the $N \times d_{\text{VLM}}$ VLM activation matrix $\mathbf{Y}$, both centered across instances, it is defined as:

$$\text{CKA}(\mathbf{X}, \mathbf{Y}) = \frac{\|\mathbf{Y}^\top \mathbf{X}\|_F^2}{\|\mathbf{X}^\top \mathbf{X}\|_F \, \|\mathbf{Y}^\top \mathbf{Y}\|_F}. \tag{6}$$

Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008) computes the Spearman correlation between the pairwise representational dissimilarity matrices (RDMs) of the two systems, capturing whether both models organize individual problems in a consistent way relative to one another, independently of the geometry of their individual representation spaces. Together, CKA provides a global measure of how well the two activation spaces align, while RSA captures whether the two systems impose a consistent relational structure on the set of problem instances.

4 Results

Datasets and connectivity matrices. We evaluate on RAVEN-10000 (Zhang et al., 2019), comprising seven task types of increasing compositional complexity: **Center**, **2×2Grid**, **3×3Grid**, **L-R** (Left-Right), **U-D** (Up-Down), **O-IC** (Out-In Center), and **O-IG** (Out-In Grid), each split into 6,000 / 2,000 / 2,000 train / validation / test samples. Full descriptions are in Appendix C.

Brain connectivity matrices are derived from the publicly available ABIDE resting-state fMRI dataset[1] ($N = 476$ typically developing controls (TD) and $N = 408$ ASD individuals). Group-level connectivity matrices are obtained by aggregating subject-level networks separately for each cohort. Visualizations of these matrices are provided in Appendix C. We use functional connectivity (a statistical dependency measure derived from resting-state BOLD correlations) throughout, and we do not equate it with anatomical/structural wiring; we return to this distinction in §5. A random connectivity matrix with matched spectral radius and sparsity serves as the control baseline. Extended analyses using individual-level connectomes are reported in Appendix L. For alignment analysis, the reservoir and two vision-language models, Qwen2-VL-2B (Wang et al., 2024) and Phi-3.5-Vision (Abdin et al., 2024), solve 2,000 matched test instances per task under identical candidate ordering. All experiments are conducted over five random seeds; performance results are reported as mean ± SD. Additional accuracy baselines, alignment-sensitivity results for three further VLMs, and robustness controls are reported in Appendices E and M.

[1] https://preprocessed-connectomes-project.org/abide/download.html

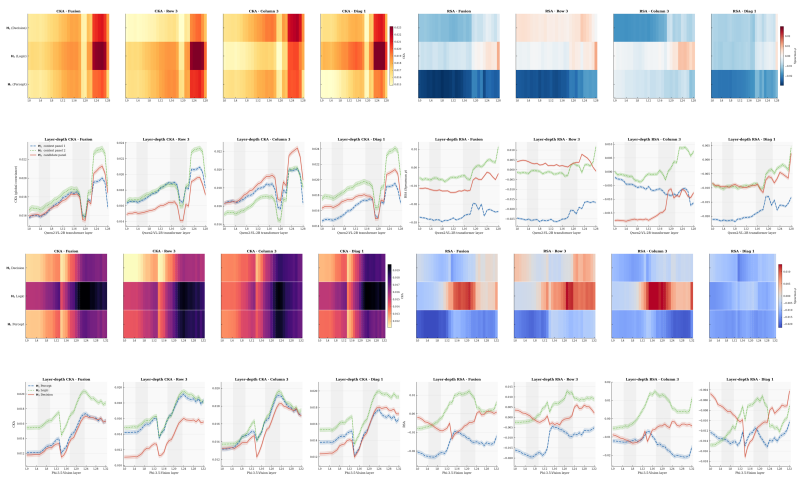


Figure 2: CKA and RSA alignment between CWR states and VLM layers: Qwen2-VL-2B (top) and Phi-3.5-vision (bottom), on CENTER task. Fusion mode + reasoning routes: Row3, Column3, Diag1.

4.1 DEPTH-ORDERED ALIGNMENT BETWEEN CONNECTOME-WIRED RESERVOIR AND VLM

Alignment structure across reasoning steps and model depth. Figure 2 shows CKA and RSA alignment between reservoir steps and Qwen2-VL-2B (top) and Phi-3.5-vision (bottom), using the TD connectome ($\alpha = 0.2$, Center task), evaluating *Hypothesis 1*. **CKA** reveals a two-phase profile with a stage-ordering reversal across depth. In early-to-middle layers (L_0–L_{15}), $\mathbf{H}_2$ (intermediate integration) maintains the highest alignment while $\mathbf{H}_3$ (terminal integration) is consistently the lowest, reflecting the priority of context-rule abstraction in the VLM's intermediate representations. A sharp transition at L_{16}–L_{19} collapses alignment across all three stages; after this, all steps recover and the ordering reverses: $\mathbf{H}_2$ reaches the highest peak in deep layers (L_{22}–L_{24}), followed closely by $\mathbf{H}_3$, while $\mathbf{H}_1$ (early integration) falls back. This reversal is not monotonic but reveals a depth-structured dissociation: the intermediate-integration ($\mathbf{H}_2$) and terminal-integration ($\mathbf{H}_3$) stages both concentrate alignment in the deepest VLM layers, while the early-integration stage decouples. **RSA** reveals a complementary and structurally cleaner pattern that holds continuously across all 29 layers. $\mathbf{H}_1$ maintains a strongly negative correlation throughout, deepening further in mid-to-deep layers, while $\mathbf{H}_3$ begins negative and rises progressively, crossing zero around L_{20} and reaching positive values in deep layers (L_{21}–L_{28}). $\mathbf{H}_2$ occupies a stable intermediate band: consistently more negative than $\mathbf{H}_3$ but less so than $\mathbf{H}_1$. This three-band ordered gradient $\mathbf{H}_1 < \mathbf{H}_2 < \mathbf{H}_3$ holds globally from L_0 to L_{28}, confirming that the step-wise relational ordering is a property of the full alignment structure rather than of deep layers alone. Crucially, both patterns replicate with Phi-3.5-vision model: the same CKA two-phase reversal and the same three-band RSA gradient ($\mathbf{H}_1 < \mathbf{H}_2 < \mathbf{H}_3$) emerge across a qualitatively different VLM architecture, confirming that the alignment structure is not model-specific but reflects a general property of learned abstract visual reasoning representations.

Together, CKA captures depth-structured covariance concentration with a stage-ordering reversal at the early-to-deep transition, while RSA captures a globally consistent relational ordering that strengthens in deep layers. Both measures, across both VLMs, directly confirm *Hypothesis 1*: alignment between the connectome-wired reservoir and trained vision-language models is systematically structured across reasoning step and model depth. Both patterns replicate across all three spatial routes and all seven task types (Appendix H).

Representational alignment is selective for correct reasoning. Since context panels are shared across all candidates, $\mathbf{H}_1$ and $\mathbf{H}_2$ are numerically identical between correct and incorrect trials by construction, so any alignment difference originates at $\mathbf{H}_3$, giving a clean, unconfounded test of *Hypothesis 2*. Figure 3 shows CKA and RSA heatmaps for the Center task under correct vs. error candidate integration, for both Qwen2-VL-2B and Phi-3.5-Vision. Under **CKA**, correct trials show a step-ordered hierarchy with $\mathbf{H}_3$ (terminal integration) reaching the highest deep-layer alignment

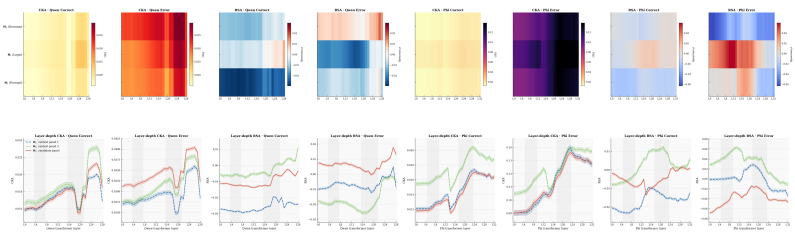


Figure 3: Correctness-conditioned CWR–VLM alignment (**Center** task). CKA and RSA heatmaps (top) and layer-depth profiles (bottom) between CWR steps (H_1–H_3) and layers of Qwen2-VL-2B (left) and Phi-3.5-Vision (right), under correct vs. error candidate integration.

Table 1: Accuracy (%) and Cohen's d effect sizes: CWR-TD and CWR-ASD vs. CWR-random. TD: typically developing. ASD: autism spectrum disorder. Cohen's d is interpreted using the conventional thresholds 0.2 (small), 0.5 (moderate), 0.8 (large), and 1.2 (very large).

	Ctr	**2×2G**	**3×3G**	**L-R**	**U-D**	**O-IC**	**O-IG**
CWR-TD	76.44	41.52	45.32	69.87	75.30	67.76	57.50
CWR-ASD	77.81	42.30	45.49	70.18	74.62	67.67	56.38
CWR-random	73.98	40.33	41.10	63.82	70.73	62.91	48.73
Cohen's d vs. CWR-random							
TD d	0.82	0.33	**1.89**	0.72	0.79	1.00	**1.93**
ASD d	**1.33**	0.60	**1.40**	0.67	0.66	1.04	**1.55**

and $\mathbf{H}_1$ (early integration) lowest; under error, this collapses, with $\mathbf{H}_1$ becoming dominant instead, indicating the reservoir's output regresses toward early perceptual encoding. **RSA** shows a sharper dissociation: on correct trials the three-band $\mathbf{H}_1 < \mathbf{H}_2 < \mathbf{H}_3$ gradient of Figure 2 is preserved, while on error trials it inverts, with $\mathbf{H}_3$ becoming strongly negative and $\mathbf{H}_2$ (intermediate integration) the most positive stage – under incorrect reasoning the reservoir fails to consolidate relational structure at the terminal-integration stage, leaving it stranded at an intermediate level of abstraction. This dissociation replicates across both VLMs and all seven tasks (Appendix I), supporting *Hypothesis 2*. Terminal-stage CKA is higher on correct than incorrect trials in 6 of 7 tasks for Qwen2-VL-2B (mean difference $+0.020$), and terminal RSA shows the same pattern for both Qwen2-VL-2B ($+0.010$) and Phi-3.5-Vision ($+0.013$).

Role of connectivity structure and temporal integration. Figure 4 isolates the contributions of biological wiring and temporal memory via ablation on the *Center* task; results for all remaining tasks are in Appendix J. Under biological connectivity, CKA peaks at $\mathbf{H}_3$ (terminal integration), is secondary for $\mathbf{H}_2$ (intermediate integration), and lowest for $\mathbf{H}_1$ (early integration); RSA mirrors this with a clean gradient, negative for $\mathbf{H}_1$, near-zero for $\mathbf{H}_2$, positive for $\mathbf{H}_3$, tracking the early-to-terminal integration flow. *Random connectivity* preserves a residual late-layer CKA peak but collapses the RSA gradient into an undifferentiated pattern. Table 1 quantifies this: accuracy remains comparable, but the advantage over random is geometry-specific, TD gains most on global tasks (O-IG, $d = 1.93$), ASD on local tasks (Ctr, $d = 1.33$), confirming biological topology as a *task-geometry-specific* inductive bias and supporting *Hypothesis 3*.

Removing temporal integration ($\alpha = 1.0$) produces a distinct failure mode: CKA inflates broadly while RSA collapses into an undifferentiated negative pattern, losing the positive $\mathbf{H}_3$ alignment needed for terminal integration variables. Biological topology thus governs the *relational structure* of alignment (RSA), while temporal integration governs its *step-wise differentiation*: a double dissociation we **discover** here, which, to our knowledge, no prior work has isolated as functionally separable axes of alignment between a fixed dynamical system and a trained vision-language model. An extended leak-rate sweep across five VLM architectures replicates this dissociation (Appendix F). A complementary model-side ablation, replacing the visual backbone with a null Qwen and a ViT-B encoder, confirms it is not an artifact of encoder choice (Appendix J.1).

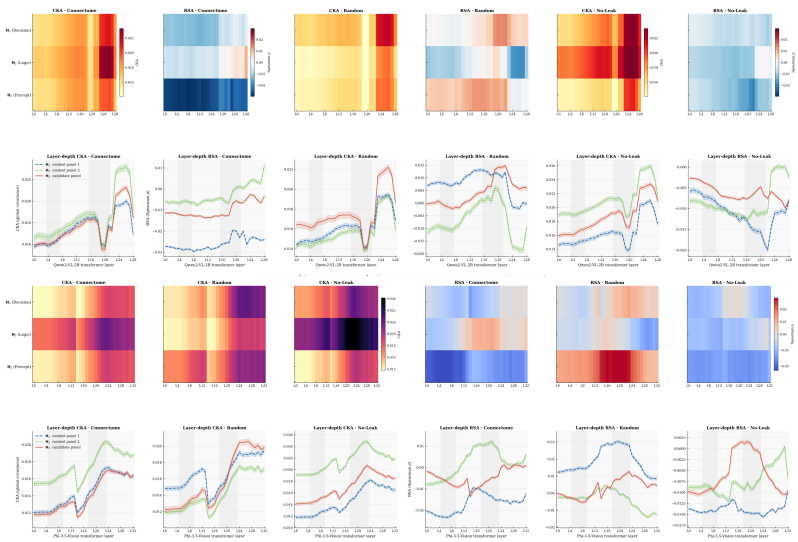


Figure 4: Ablation of connectivity structure and temporal integration. Alignment of CWR with Qwen2-VL-2B (top) and with Phi-3.5-Vision (bottom) under three conditions (Connectome, Random, no-leak).

The advantage is not explained by generic graph structure. A uniform-random control leaves open whether any sufficiently *structured* graph, not specifically the empirical one, would reproduce the same advantage. We therefore instantiate the reservoir with alternative connectivity matrices isolating specific properties of the empirical connectome: degree-preserving Maslov–Sneppen rewiring, ROI-permuted connectivity (preserving the empirical connectome's values and spectrum but destroying their regional arrangement), a stochastic block model, and a Watts–Strogatz graph (macro-averaged accuracies and per-VLM alignment ratios in Appendix M, Table 13). On *accuracy*, CWR is comparable to Maslov–Sneppen, SBM, and Watts–Strogatz (within $\sim$1 point) and exceeds ROI-permuted connectivity (61.96% vs. 57.70%); on *alignment*, the empirical connectome exceeds every one of these controls, with CKA higher in 63/63 matched comparisons against both Maslov–Sneppen and ROI-permuted connectivity. Different connectivity controls produce comparable decoding performance, while empirical FC shows distinct representational alignment characteristics. An independently constructed set of null matrices (Appendix M) reproduces this pattern on a held-out task and VLM.

Generalization beyond functional connectivity. The controls above reduce the plausibility that non-biological structured graphs explain the observed advantage. A separate question is whether the effect is specific to functional connectivity as a data modality. Repeating the analysis with an independently derived *morphological* connectome, we find the same qualitative advantage over its matched random control, though weaker and with lower absolute accuracy than the functional connectome used throughout the paper (Appendix E).

4.2 Abstract Visual Reasoning Performance of CWR

Decoding performance under connectivity controls. Under a unified framework in which ESN (Jaeger, 2001), LSM (Maass et al., 2002), a fractal-topology reservoir (Appendix D), and CWR share identical hyperparameters and readout architecture, with reservoir connectivity as the sole variable, CWR outperforms all vectorized baselines on every task, with effect sizes ranging from moderate ($d \approx 0.7$) to very large ($d > 3$, up to $d = 13.63$ on Center vs. LSM; full table and statistics in Appendix K.1). Replacing the connectome with a random matrix of matched spectral radius and density reduces accuracy by $\sim$4 points on average, with the gap widening on tasks demanding more complex relational integration. This grounds the representational-alignment analyses of §4.1 in a system that solves the task meaningfully rather than degenerately, and directly supports *Hypothesis 3*. CWR-TD and CWR-ASD perform comparably across tasks (Appendix K.2).

Table 2: CWR accuracy alongside all five evaluated VLM baselines (task-level context, not a performance comparison; see text). Qwen2-VL-2B and Phi-3.5-Vision are used in the core alignment analysis of §4.1; the remaining three extend the accuracy and sensitivity comparisons (Appendix E).

Method	Ctr	2×2G	3×3G	L-R	U-D	O-IC	O-IG
CWR-TD	76.44 ± 3.05	41.52 ± 3.93	45.32 ± 2.16	69.87 ± 8.39	75.30 ± 6.82	67.76 ± 5.61	57.50 ± 3.39
CWR-ASD	77.81 ± 2.90	42.30 ± 3.40	45.49 ± 3.52	70.18 ± 10.05	74.62 ± 6.90	67.67 ± 5.28	56.38 ± 4.15
Qwen2-VL-2B (Wang et al., 2024)	67.64 ± 1.93	41.67 ± 1.05	43.36 ± 0.55	44.57 ± 1.20	44.85 ± 1.66	47.88 ± 1.04	39.86 ± 1.80
InternVL2-2B	72.03 ± 7.67	50.46 ± 1.27	52.64 ± 1.39	58.10 ± 1.90	71.93 ± 1.60	66.67 ± 1.01	62.50 ± 1.09
Qwen2.5-VL-3B	86.14 ± 1.10	57.10 ± 1.60	47.47 ± 1.96	89.72 ± 3.94	45.01 ± 3.42	82.35 ± 1.49	75.98 ± 2.16
Phi-3.5-Vision (Abdin et al., 2024)	90.63 ± 0.83	56.23 ± 1.33	60.02 ± 1.02	87.22 ± 3.18	89.73 ± 0.72	76.73 ± 1.90	76.92 ± 2.88
PaliGemma2-3B	98.63 ± 0.35	77.55 ± 1.08	81.25 ± 0.39	97.59 ± 0.33	98.26 ± 0.14	97.22 ± 0.33	94.04 ± 0.29

Structural inductive bias vs. scale and pretraining. Table 2 reports accuracy for CWR alongside all five VLMs evaluated in this study. These comparisons provide task-level context rather than a performance competition: our primary objective is to characterize whether representations from fixed connectome-derived dynamics align systematically with those of trained VLMs (§4.1), not to match or outperform large pretrained models on accuracy. Despite no visual pretraining or language supervision, CWR exceeds Qwen2-VL-2B and InternVL2-2B on several tasks, with its largest margins on abstract spatial tasks (L-R, U-D); it is not accuracy-competitive with the strongest architectures tested, Phi-3.5-Vision and PaliGemma2-3B, the latter exceeding 94% on every task. We do not read these comparisons as positioning CWR as an accuracy-competitive alternative to modern, large-scale VLMs; their role is to ground the alignment analysis in a system that solves the task non-trivially, without visual pretraining, and at a fraction of the training cost (8–10 minutes on a consumer-grade GPU vs. 5–10 hours of fine-tuning; Appendix D). The NKP-rank and leak-rate sensitivity results reported above hold across all five architectures, and a candidate-order shuffling control found no statistically detectable sensitivity to option order under this test (relative fluctuation $\leq 3.07\%$ across all seven tasks; Appendix K.3), indicating that performance is not attributable to positional artifacts (Mitchell, 2021).

5 Discussion

Brain wiring and VLM representations. Our central finding is that a reservoir with fixed empirical functional connectivity produces representations that align with trained VLMs without updating its recurrent weights. This alignment is ordered across reasoning stage and VLM depth, can be stronger for correct than incorrect candidates in several model–metric combinations, and dissociates into two separable contributors: connectivity organization shapes the *relational structure* of alignment, while temporal integration governs its *step-wise differentiation*. Neither factor alone reproduces the full pattern, and the dissociation holds across five VLM architectures and is not attributable to encoder choice (Appendices F, J.1). Structured, non-biological controls further reduce the plausibility that density, degree sequence, and community structure alone explain the alignment (Appendix M). Throughout, alignment is measured between CWR and VLM representations; no neural recordings are used. The TD and ASD connectomes both outperform matched random wiring in the evaluated comparisons, although the relative advantage differs across tasks, so these differences should not be interpreted as differences in cognitive ability or diagnostic status (Just et al., 2004; Geschwind & Levitt, 2007; Belmonte et al., 2004).

Limitations and future directions. The structured-null controls (Table 13) reduce the plausibility that density, degree sequence, and community structure alone explain for the alignment advantage, but they do not yet identify which specific property of the empirical connectome's regional arrangement is responsible; resolving this is the clearest next step, and would come from testing the empirical connectome against a distribution of independently generated spectrum-matched synthetic operators rather than a single instance. Depth-ordered alignment is established for all five evaluated VLMs (Appendix H), and a natural extension is testing whether the same structure holds for larger, more capable VLM families as they become available. Beyond these, we see value in moving past RAVEN to broader relational-reasoning tasks and naturalistic visual inputs, and in eventually relating connectome-derived reservoir states to measured neural activity directly, since the present study compares reservoir states only to trained VLM representations throughout.

REFERENCES

Marah Abdin, Sam Adé Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Singh Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. doi: 10.48550/arXiv.2404.14219.

David Barrett, Felix Hill, Adam Santoro, Ari S. Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 511–520. PMLR, 2018. URL https://proceedings.mlr.press/v80/barrett18a.html.

Matthew K. Belmonte, Gregory Allen, Andrea Beckel-Mitchener, Lisa M. Boulanger, Ruth A. Carper, and Sara J. Webb. Autism and abnormal development of brain connectivity. *Journal of Neuroscience*, 24(42):9228–9231, 2004. doi: 10.1523/JNEUROSCI.3340-04.2004.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Deep language algorithms predict semantic comprehension from brain activity. *Scientific Reports*, 12(1):16327, 2022.

Fabrizio Damicelli, Claus C Hilgetag, and Alexandros Goulas. Brain connectivity meets reservoir computing. *PLoS Computational Biology*, 18(11):e1010639, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Michael Eickenberg, Alexandre Gramfort, G. Varoquaux, and B. Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 2017. doi: 10.1016/j.neuroimage.2016.10.001.

Daniel H. Geschwind and Pat Levitt. Autism spectrum disorders: developmental disconnection syndromes. *Current Opinion in Neurobiology*, 17(1):103–111, 2007. doi: 10.1016/j.conb.2007.01.009.

Eghbal A. Hosseini, Colton Casto, Noga Zaslavsky, Colin Conwell, Mark Richardson, and Evelina Fedorenko. Universality of representation in biological and artificial neural networks. *bioRxiv*, 2024. doi: 10.1101/2024.12.26.629294.

Sheng Hu, Yuqing Ma, Xianglong Liu, Yanlu Wei, and Shihao Bai. Stratified rule-aware network for abstract visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1567–1574, 2021. doi: 10.1609/aaai.v35i2.16248.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning*. PMLR, 2024.

Herbert Jaeger. The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, 148(34):13, 2001.

Marcel Adam Just, Vladimir L Cherkassky, Timothy A Keller, and Nancy J Minshew. Cortical activation and synchronization during sentence comprehension in high-functioning autism: evidence of underconnectivity. *Brain*, 127(8):1811–1821, 2004.

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.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 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*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Mantas Lukoševičius and Herbert Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149, 2009. doi: 10.1016/j.cosrev.2009.03.005.

Wolfgang Maass, Thomas Natschläger, and Henry Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural computation*, 14 (11):2531–2560, 2002.

Melanie Mitchell. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101, 2021. doi: 10.1111/nyas.14588.

Alex Murphy, Joel Zylberberg, and Alona Fyshe. Correcting biased centered kernel alignment measures in biological and artificial neural networks. In *ICLR 2024 Workshop on Representational Alignment (Re-Align)*, 2024. doi: 10.48550/arXiv.2405.01012. arXiv:2405.01012.

Kohei Nakajima and Ingo Fischer. *Reservoir computing*. Springer, 2021.

Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. *arXiv preprint arXiv:2010.15327*, 2020.

Joséphine Raugel, Stéphane d'Ascoli, Jérémy Rapin, Valentin Wyart, and Jean-Rémi King. Scaling and context steer llms along the same computational path as the human brain. *arXiv preprint arXiv:2512.01591*, 2025.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

Guobin Shen, Dongcheng Zhao, Yiting Dong, Qian Zhang, and Yi Zeng. Alignment between brains and ai: Evidence for convergent evolution across modalities, scales and training trajectories. *arXiv preprint arXiv:2507.01966*, 2025. doi: 10.48550/arXiv.2507.01966.

Mayssa Soussia, Mohamed Ali Mahjoub, and Islem Rekik. Multi-sensory cognitive computing for learning population-level brain connectivity. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 521–531. Springer, 2025.

Laura E Suárez, Blake A Richards, Guillaume Lajoie, and Bratislav Misic. Learning function from structure in neuromorphic networks. *Nature Machine Intelligence*, 3(9):771–786, 2021.

Laura E Suárez, Agoston Mihalik, Filip Milisav, Kenji Marshall, Mingze Li, Petra E Vértes, Guillaume Lajoie, and Bratislav Misic. Connectome-based reservoir computing with the conn2res toolbox. *Nature communications*, 15(1):656, 2024.

Charles F Van Loan and Nikos Pitsianis. Approximation with kronecker products. In *Linear algebra for large scale and real-time applications*, pp. 293–314. Springer, 1993.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541, 2023.

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

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.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5317–5327, 2019.

A Appendix

B Methodological notes

A note on gradient. A central design property of CWR is the strict separation between the fixed dynamical system and the trained readout. The bilinear operators $\mathbf{W}_L$ and $\mathbf{W}_R$ are derived from the brain connectivity matrix and are *never updated*: no gradient flows through the reservoir state update (Eq. 2) at any point during training. Formally, all reservoir parameters are wrapped in a `torch.no_grad()` context during the forward pass, and their `requires_grad` flag is set to `False` at initialization. The input projection matrices $\mathbf{P}_L$ and $\mathbf{P}_R$ are likewise fixed. The only components that receive gradient updates are (i) the semantic latent projection $\mathbf{W}_{\text{proj}}$ that maps the 768-dimensional temporal feature vector to the D-dimensional latent space, (ii) the three attention stages, (iii) the MLP relational fusion networks, and (iv) the gating network. This isolation means that the representational geometry induced in $\{\mathbf{H}_1, \mathbf{H}_2, \mathbf{H}_3\}$ is determined *entirely* by the brain connectivity structure, not by any downstream learning objective. Any alignment between these states and a trained VLM therefore cannot be attributed to gradient-based optimization. We verified this directly (not only architecturally): across all seven tasks and five seeds, stored reservoir states are numerically identical before and after readout training, with a maximum difference of zero.

A note on alignment measures. The CKA–RSA dissociation observed for random-weight reservoirs has implications beyond this paper. High-dimensional random projections produce spuriously elevated CKA even between unrelated representations, a well-documented sensitivity of biased CKA to the feature-to-sample ratio (Nguyen et al., 2020; Murphy et al., 2024), an artifact that RSA, which operates on pairwise relational geometry, exposes cleanly. Any alignment claim for high-dimensional fixed-weight systems that rests on CKA alone should therefore be accompanied by a geometry-sensitive measure; the two measures together separate genuine structured correspondence from dimensional coincidence. We also directly checked whether our central CKA comparisons are artifacts of the feature-to-sample ratio or of high dimensionality: recomputing the empirical-vs.-Maslov–Sneppen CKA difference under instance subsampling (250/500/1,000 instances), under a shared 128-dimensional PCA reduction of both representations, and under a debiased CKA estimator preserved the sign of the empirical-connectome advantage for both Qwen2-VL-2B and Phi-3.5-Vision in every case, though the debiased estimate was smaller in magnitude than the raw estimate and some task-level confidence intervals crossed zero under debiasing (full detail in Appendix M). We report this as a caveat: the direction of the empirical-connectome advantage is robust to these estimator choices, but its precise magnitude is estimator-dependent.

B.1 Fractal reservoir variant

The fractal reservoir is a multi-timescale extension of the standard ESN in which the state space is partitioned into S non-overlapping subpopulations, each assigned a leak rate following the harmonic progression $\alpha_i = {}^{1}/_{(i+1)}$ for $i \in \{1, \ldots, S\}$. The recurrent weight matrix is block-diagonal (no direct inter-scale connections exist) so each subpopulation evolves independently:

$$\mathbf{X}_i(t) = \big(1 - \alpha_i\big)\mathbf{X}_i(t-1) + \alpha_i\, f\Big(\mathbf{W}_{ii}\,\mathbf{X}_i(t-1) + \mathbf{W}_{\text{in},i}\,\mathbf{u}(t)\Big). \tag{7}$$

The harmonic progression produces memory time constants $\tau_i \propto (i+1)$ uniformly spaced on a logarithmic scale, engineering a self-similar temporal frequency spectrum that encodes the same input simultaneously at multiple levels of temporal abstraction, from fast-adapting to slow-integrating scales, without any learned recurrent parameters.

The rationale for including this baseline is to test whether structured multi-timescale dynamics alone, independent of biological wiring, are sufficient to reproduce the representational alignment observed with CWR.

B.2 Logic descriptor construction

High-order temporal feature extraction. For each reasoning path $r \in \{\text{row}, \text{col}, \text{diag}\}$, the reservoir produces a trajectory of three states $\{\mathbf{H}_1, \mathbf{H}_2, \mathbf{H}_3\} \subset \mathbb{R}^{d \times d}$ where $d = 16$. These are

flattened and concatenated into a 768-dimensional multi-scale representation that jointly captures the final reasoning state, the late-stage logical transition, and the early-stage rule initiation:

$$\mathbf{v} = \big[\mathrm{vec}(\mathbf{H}_3) \parallel \mathrm{vec}(\mathbf{H}_3 - \mathbf{H}_2) \parallel \mathrm{vec}(\mathbf{H}_2 - \mathbf{H}_1)\big] \in \mathbb{R}^{768}, \tag{8}$$

where $\mathrm{vec}(\mathbf{H}_3)$ encodes global logical accumulation across the full sequence, $\mathrm{vec}(\mathbf{H}_3 - \mathbf{H}_2)$ (late delta) captures rule convergence from the second to third panel, and $\mathrm{vec}(\mathbf{H}_2 - \mathbf{H}_1)$ (early delta) captures rule initiation from the first to second panel. This representation is projected into a $D = 112$ semantic latent space via a learnable GELU projection:

$$\mathbf{z} = \mathrm{GELU}(\mathbf{W}_{\mathrm{proj}}\,\mathbf{v} + \mathbf{b}_{\mathrm{proj}}) \in \mathbb{R}^D. \tag{9}$$

The GELU nonlinearity suppresses reservoir noise and retains only salient logical features for subsequent relational modeling.

Differential relational modeling. In latent space, for each path r, the framework extracts a *Rule Vector* $\mathbf{r}_r$ and a candidate-specific *Candidate Vector* $\mathbf{c}_r^{(k)}$:

$$\mathbf{r}_r = \mathbf{z}_{r,2} - \mathbf{z}_{r,1}, \qquad \mathbf{c}_r^{(k)} = \mathbf{z}_{r,3}^{(k)} - \mathbf{z}_{r,2}, \qquad \mathbf{e}_r^{(k)} = \mathbf{c}_r^{(k)} - \mathbf{r}_r, \tag{10}$$

where indices 0, 1, 2 correspond to the early-delta, late-delta, and final-state components of $\mathbf{v}$. For the diagonal route, a dedicated step-wise analogy decomposes reasoning into a corner-to-center step (anti-diagonal rule) and a center-to-candidate step (main-diagonal candidate):

$$\mathbf{r}_{\mathrm{diag}} = \mathbf{z}_{\text{anti-diag}}^{(H_2-H_1)}, \qquad \mathbf{c}_{\mathrm{diag}}^{(k)} = \mathbf{z}_{\text{main-diag},k}^{(H_3-H_2)}. \tag{11}$$

The concatenated rule–candidate pair $[\mathbf{r}_r \| \mathbf{c}_r^{(k)}]$ is fused via a path-specific MLP: $\mathbf{f}_{r,k} = \mathrm{MLP}_{\mathrm{rel}}([\mathbf{r}_r \| \mathbf{c}_r^{(k)}]) \in \mathbb{R}^D$. The three route-wise feature blocks are stacked into a single Relational Reasoning Tensor $\mathcal{T}_{\mathrm{rel}} \in \mathbb{R}^{B\times K\times 3\times D}$.

Hierarchical attention and decision. The relational reasoning tensor $\mathcal{T}_{\mathrm{rel}} \in \mathbb{R}^{B\times K\times 3\times D}$ is processed by three successive attention stages, each resolving a different aspect of the reasoning problem.

Stage 1 — Cross-route self-attention. For each candidate k, the three route-wise feature vectors (row, column, diagonal) attend to one another. This enforces *internal consistency*: if a candidate is correct, its induced transformation should be coherent across all three spatial directions simultaneously.

Stage 2 — Rule-conditioned cross-attention. The Logical Benchmark Tensor $\mathcal{R} \in \mathbb{R}^{B\times 3\times D}$, formed by stacking the rule vectors $\mathbf{r}_r$ extracted from context paths, serves as Key and Value. Each candidate's feature vector serves as Query. This stage scores how well the transformation induced by candidate k along each route matches the relational rule inferred from context — the core analogical reasoning step.

Stage 3 — Cross-candidate self-attention. All $K = 8$ candidates attend jointly to one another, enabling relative comparison rather than independent absolute scoring. This allows the model to identify which candidate stands out as most rule-consistent given the full answer set.

After the three stages, route-wise scalar scores $s_{k,r} \in \mathbb{R}$ are augmented with explicit geometric scalars — cosine similarity, norm ratio, and scaling error between $\mathbf{r}_r$ and $\mathbf{c}_r^{(k)}$ — capturing fine-grained relational alignment beyond what attention encodes. A Gated Mixture-of-Experts head then aggregates route-wise evidence into a final score:

$$s_k^{\mathrm{final}} = \sum_{r\in\{\mathrm{row,col,diag}\}} s_{k,r}\cdot w_r, \qquad w_r \in [0,1],\ \sum_r w_r = 1, \tag{12}$$

where the gate weights $\mathbf{w}$ are produced by a small network conditioned on the global feature distribution, allowing the model to up-weight whichever spatial route carries the most discriminative signal for each problem instance.

Tensor-level construction.

```
states: [B,29,768]
768 = vec(H3)[256] || vec(H3-H2)[256] || vec(H2-H1)[256]
indices 0..4   = row1, row2, col1, col2, anti-diagonal contexts
indices 5:13   = 8 row candidates
indices 13:21  = 8 column candidates
indices 21:29  = 8 diagonal candidates

h = state_proj(states)                                  # [B,29,D], D=112
rule_row = h[:,1]-h[:,0];  rule_col = h[:,3]-h[:,2]  # [B,D]
diff_row = h[:,5:13]  - h[:,1,None]                   # [B,8,D]
diff_col = h[:,13:21] - h[:,3,None]                   # [B,8,D]
rule_diag = rule_diag_proj(states[:,4,512:768])       # [B,D]
diff_diag = diff_diag_proj(states[:,21:29,256:512])  # [B,8,D]

feat_r = diff_proj([diff_r || repeat(rule_r,8)])  # [B,8,D], per route
cand = stack(feat_row, feat_col, feat_diag, axis=2)   # [B,8,3,D]
```

Cross-route self-attention (Stage 1) runs over sequence length 3 after reshaping to $[B \times 8, 3, D]$. Rule-conditioned cross-attention (Stage 2) uses candidate-route features as queries and the three rule vectors as keys and values over the route axis. Cross-candidate self-attention (Stage 3) then runs over the eight-candidate axis, separately within each route.

C Dataset and connectivity details

RAVEN benchmark. We evaluate on RAVEN-10000 (Zhang et al., 2019), a large-scale abstract visual reasoning benchmark comprising seven task configurations of increasing compositional complexity. Each configuration presents a 3×3 matrix governed by a hidden relational rule; one entry is missing and must be selected from eight candidates. The seven configurations are: **Center**, **2×2Grid**, **3×3Grid**, **L-R** (Left-Right), **U-D** (Up-Down), **O-IC** (Out-In Center), and **O-IG** (Out-In Grid). Each is split into 6,000 / 2,000 / 2,000 train / validation / test samples. For the representational alignment analysis, 2,000 matched test instances per task are used, with candidate ordering held identical across CWR and both VLMs to ensure a controlled comparison.

Figure 5 illustrates all seven configurations. For each, the upper panel shows the 3×3 context matrix with the missing entry marked ?, and the lower panel shows the eight answer candidates arranged in a 2×4 grid. **(a) Center**: a single object occupies the central cell; size and shade vary systematically across rows. **(b) 2×2Grid**: four objects arranged in a 2×2 layout; object number and type change jointly across rows and columns. **(c) 3×3Grid**: nine objects fill all cells; color and position shift systematically, requiring integration of multiple co-varying attributes. **(d) L-R** (Left-Right): two objects placed left and right of center; attributes mirror or progress across columns. **(e) U-D** (Up-Down): two objects placed above and below center; analogous vertical relational structure to (d). **(f) O-IC** (Out-In Center): a frame surrounds a central object; two attributes co-vary between the outer frame and inner object. **(g) O-IG** (Out-In Grid): a frame surrounds a 2×2 inner grid; the most compositionally complex configuration, requiring simultaneous reasoning over inner and outer relational structure.

Functional connectivity matrices (ABIDE I). Functional connectivity matrices are derived from the Autism Brain Imaging Data Exchange I (ABIDE I), accessed via the Preprocessed Connectomes Project.[2] After quality-control filtering, the cohort comprises $N=476$ typically developing (TD) and $N=408$ ASD subjects (884 total). Preprocessing follows the CPAC pipeline with band-pass filtering (0.01–0.1 Hz), motion scrubbing, and nuisance regression (6 motion parameters, global signal, white matter and CSF signals). ROI-level BOLD time-series are extracted using the Harvard-Oxford Atlas ($R=111$ regions), yielding subject-level functional connectivity matrices via Pearson correlation of the ROI BOLD time-series. Group-level connectivity matrices are obtained by mean aggregation across subjects within each cohort (TD and ASD separately), producing two 111×111 symmetric matrices reflecting population-level functional connectivity strength across ROI pairs (range ≈-0.33 to 0.71; Figure 6). The weakest 10% of connections by absolute weight are pruned (90% retention), a threshold determined by grid search as the point at which biological connectivity most clearly

[2] http://preprocessed-connectomes-project.org/abide/

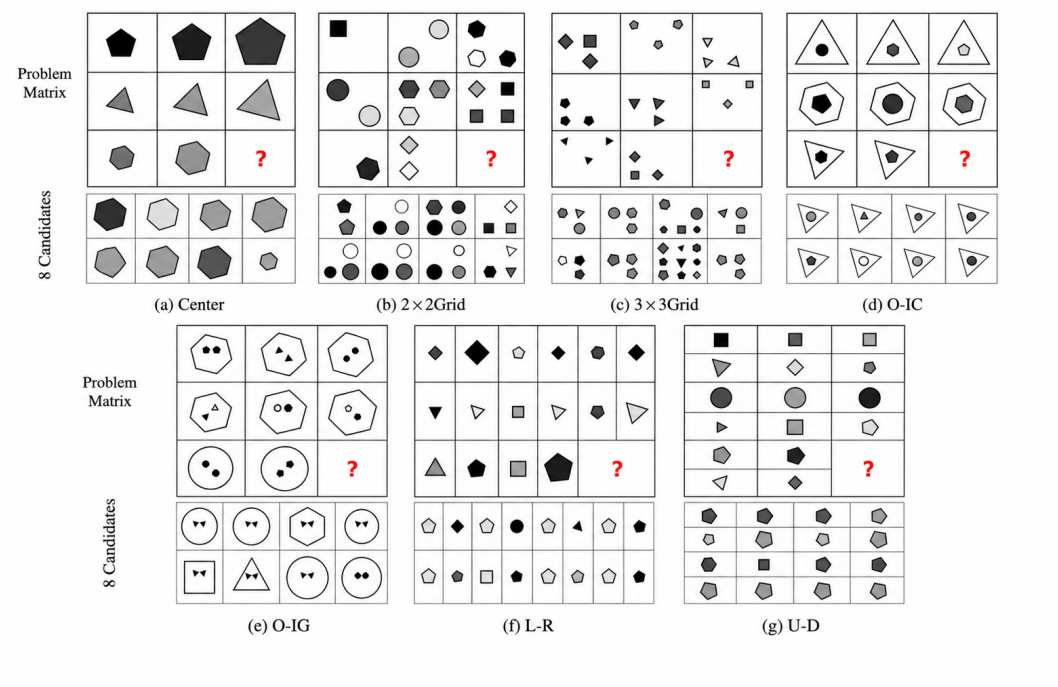
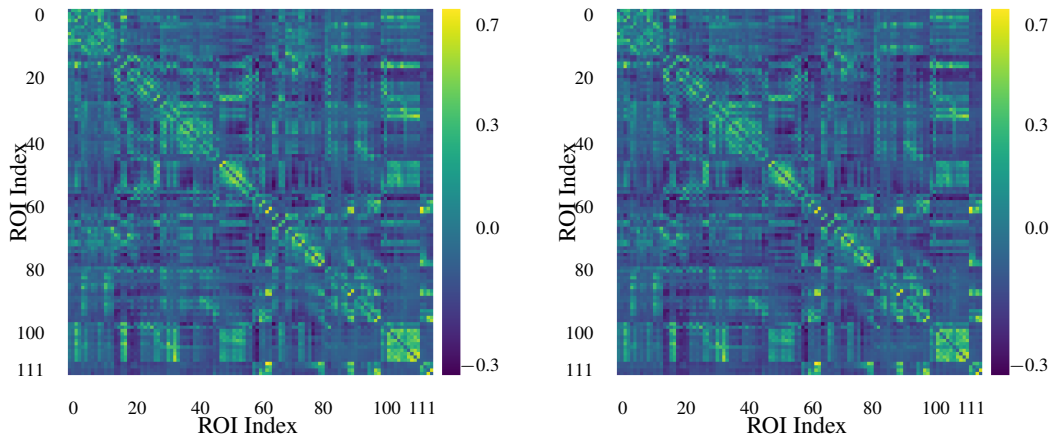


Figure 5: All seven RAVEN task configurations. For each configuration the upper panel is the 3×3 context matrix (missing entry marked ?) and the lower panel shows the eight answer candidates in a 2×4 grid. **(a) Center**: a single object in the central cell; size and shade vary across rows. **(b) 2×2Grid**: four objects in a 2×2 arrangement; number and type change jointly. **(c) 3×3Grid**: nine objects fill all cells; colour and position shift systematically. **(d) L-R** (Left-Right): two objects placed left and right of centre; attributes mirror or progress across columns. **(e) U-D** (Up-Down): two objects placed above and below centre; analogous vertical structure to (d). **(f) O-IC** (Out-In Center): a frame surrounds a central object; two attributes co-vary. **(g) O-IG** (Out-In Grid): a frame surrounds a 2×2 inner grid; the most compositionally complex configuration.

outperforms matched random connectivity. Each group-level matrix is then upsampled to 256×256 via linear interpolation, followed by absolute-value transformation, random sign assignment, and normalization by the maximum absolute value. The resulting matrix is factorized via Rank-1 NKP decomposition into bilinear operators $\mathbf{W}_L, \mathbf{W}_R \in \mathbb{R}^{16\times16}$ for use as fixed recurrent operators in CWR (§3.2). Subject-level matrices are used directly for individual-level analyses (Appendix L). We separately tested how much of this preprocessing pipeline drives the reported results; the interpolation step, the sign-assignment step, and the choice of rank are each addressed by a dedicated control in Appendix M.

D Implementation and Hyperparameter Details

Bilinear NKP decomposition. The sparsified $R\times R$ connectivity matrix (functional: 256×256 after upsampling) is linearly interpolated to $d^2\times d^2$ ($d=16$, so 256×256) using `scipy.ndimage.zoom`. The NKP problem is solved via the rearrangement operator (Van Loan & Pitsianis, 1993): reshape to a $d^2\times d^2$ block matrix, compute the leading rank-1 SVD, and reshape singular vectors to obtain $\mathbf{W}_L, \mathbf{W}_R \in \mathbb{R}^{16\times16}$. Sensitivity analysis confirmed that higher-rank approximations increase parameter count without significant gains in reasoning accuracy. The spectral radii of $\mathbf{W}_L$ and $\mathbf{W}_R$ are independently rescaled to 1.05, a value selected at the edge of stability to maintain sufficient non-linear expressivity (Suárez et al., 2021); the sweep $\rho \in \{0.8, 0.9, 1.0, 1.1\}$ confirmed this choice. The rank of the NKP decomposition (the number of Kronecker factor pairs) is distinct from the matrix rank of the reconstructed recurrent operator; see §3.2 and Appendix G.



(a) TD group-level connectivity

(b) ASD group-level connectivity

Figure 6: Group-level functional connectivity matrices for the TD (left) and ASD (right) cohorts derived from ABIDE I. Both axes index the $R = 111$ ROIs of the Harvard-Oxford Atlas. Color intensity reflects functional connectivity strength (range ≈ -0.33 to 0.71). Each matrix is subsequently resampled and factorized via NKP decomposition into fixed bilinear operators $\mathbf{W}_L$ and $\mathbf{W}_R$ for use in CWR.

Random connectivity control. The random baseline follows the identical pipeline as the biological group-level connectome to ensure a fair comparison. Specifically: (1) a 111×111 uniform random matrix with randomly assigned signs is generated; (2) it is linearly interpolated to 256×256; (3) the weakest 10% of connections are pruned (90% retention); and (4) the resulting matrix undergoes the same NKP bilinear decomposition to produce $\mathbf{W}_L$ and $\mathbf{W}_R$ of dimension 16×16. The random baseline therefore shares the same matrix dimensions, sparsity, interpolation procedure, and factorization method as the biological group-level connectome; the only difference is the absence of biological structure. This uniform-random baseline is our weakest null; Appendix M reports substantially more stringent structure-matched nulls (degree-preserving rewiring, region permutation, stochastic block models, small-world graphs).

CWR hyperparameters. Table 3 summarizes all reservoir hyperparameters and their selection criteria. The input scaling factor $\gamma = 4.0$ ensures that the projected signal carries sufficient energy to drive reservoir neurons into the non-linear operating range of the $\tanh$ activation. Resolution $d = 16$ was selected after evaluating $d \in \{8, 16, 20\}$: $d = 8$ is insufficiently expressive and $d = 20$ over-parameterize the NKP factorization without accuracy gains.

Table 3: CWR reservoir hyperparameters and selection criteria.

Parameter	Value	Sweep	Criterion
State resolution d	16	$\{8, 16, 20\}$	Validation accuracy
Leak rate α	0.2	$\{0.1, 0.2, \ldots, 1.0\}$ (Appendix F)	Accuracy + RSA at $\mathbf{H}_3$
Input scaling γ	4.0	Grid search	Non-linear activation range
Sparsity threshold	0.90	Grid search	Max biological vs. random gap
NKP rank	1	$\{1, 2, 4\}$ (Appendix G)	Accuracy / parameter trade-off

Fractal reservoir baseline. The fractal reservoir uses $S = 4$ strata of 64 neurons each, matching the 256-neuron total capacity of CWR. All strata use a *uniform* leak rate $\alpha = 0.2$, identical to the CWR setting; the only distinction from ESN is the hierarchical block-diagonal connectivity topology (detailed in Appendix B.1). Each stratum is assigned a spectral radius of 1.05 per factor, preventing instability across the cascaded hierarchy. This design ensures that the fractal baseline isolates the effect of structured hierarchical topology from any difference in temporal dynamics.

ESN and LSM baselines. Both baselines use spectral radius $1.05^2 \approx 1.10$ to match the effective dynamical activity of CWR, and the same sparsity threshold (0.90), input scaling ($\gamma = 4.0$), and readout architecture.

Readout architecture and training. The semantic latent space has dimension $D = 112$. Training uses AdamW (lr $= 3 \times 10^{-4}$, weight decay $= 2 \times 10^{-4}$), batch size 64, 100 epochs with CosineAnnealingLR (decaying to 10^{-5}). Loss weights $\lambda_r = 1.0$ (ranking) and $\lambda_c = 0.1$ (cross-route consistency) were selected by grid search. Random seeds: $\{42, 52, 62, 72, 82\}$; `seed_everything` is called at the start of every (task, condition, seed) block to guarantee full RNG isolation.

VLM fine-tuning. Both VLMs are fine-tuned via LoRA ($r = 16$, lora_alpha $= 32$, lora_dropout $= 0.05$) injected into all linear projection layers; the vision encoder and pre-trained weights are frozen. Qwen2-VL-2B: AdamW, lr $= 5 \times 10^{-5}$, 5 epochs, batch size 16, bfloat16; best validation checkpoint retained per seed. Phi-3.5-vision: AdamW, lr $= 5 \times 10^{-5}$, 3 epochs, batch size 2 with gradient accumulation $\times 4$ (effective batch 8), gradient checkpointing, bfloat16; final epoch used.

Hidden state extraction and alignment protocol. VLM hidden states are mean-pooled across the full token dimension at each layer, including both image patch tokens and text tokens, yielding a tensor of shape [Samples $\times$ Layers $\times$ Hidden_Dim]. CWR states $\mathbf{H}_{t,r} \in \mathbb{R}^{16 \times 16}$ are vectorized to $\mathbb{R}^{256}$ and averaged across the three routes (Row3, Col3, Diag1) for fusion-mode analysis. Both representations are standardized (zero mean, unit variance) across $N = 2{,}000$ instances. Alignment scores are computed per seed and then averaged across five seeds for both VLMs.

Position-bias control. For the option-shuffling experiment, the 8 candidate panels are deterministically permuted by sample index with synchronous ground-truth update, eliminating positional bias without introducing stochasticity.

Computational setup. CWR converges within 8–10 minutes per task–seed run on a consumer-grade workstation (Intel i5-11400, 16 GB RAM, NVIDIA RTX 3070 Laptop 8 GB, Windows 11). Qwen2-VL-2B fine-tuning required $\sim$5 hours and Phi-3.5-vision $\sim$9–10 hours per task–seed run (Intel Xeon Platinum 8470Q, 754 GB RAM, NVIDIA RTX 5090 32 GB, Ubuntu 22.04).

E Additional vision-language models and morphological-connectome robustness

Table 4 summarizes the five vision-language models used in this study: the two evaluated in the main representational-alignment analysis (Section 4.1), and three additional models — InternVL2-2B, Qwen2.5-VL-3B, and PaliGemma2-3B — evaluated as independent accuracy baselines and as alignment targets in the leak-rate and NKP-rank sensitivity analyses (Appendices F and G). These five models are trained, gradient-optimized systems and are not instantiated with any brain connectome; they serve strictly as external baselines against which CWR is compared.

Table 4: Vision-language models used as accuracy baselines and alignment targets.

Model	Parameters	Vision Encoder	Institution
Qwen2-VL-2B	2B	Qwen ViT	Alibaba
Phi-3.5-Vision	4.2B	CLIP	Microsoft
InternVL2-2B	2B	Intern-ViT	Shanghai AI Lab
Qwen2.5-VL-3B	3B	Qwen2.5	Alibaba
PaliGemma2-3B	3B	SigLIP	Google

Independent VLM accuracy comparison. Table 5 reports the accuracy of the three additional VLMs on all seven RAVEN tasks, to be read alongside the CWR and original-VLM results already reported in the main paper. Against InternVL2-2B, CWR-TD and CWR-ASD are ahead on Center, L-R, U-D, and O-IC, and behind on 2$\times$2Grid, 3$\times$3Grid, and O-IG. Against Qwen2.5-VL-3B, Phi-3.5-Vision, and PaliGemma2-3B, CWR is outperformed on nearly every task, with the exception of U-D,

where CWR-TD exceeds Qwen2.5-VL-3B by more than 30 points; PaliGemma2-3B in particular attains near-ceiling accuracy ($> 94\%$) on all seven tasks.

Morphological-connectome CWR: a robustness analysis. Separately from the VLM comparison, we evaluate whether CWR remains effective when its recurrent operators are derived from a morphological connectome rather than from functional connectivity. Following the same NKP factorization and reservoir pipeline described in Section 3.2, we instantiate CWR using morphological connectomes constructed from the left- and right-hemisphere cortical morphology of the TD and ASD cohorts, denoted CWR(TD-LH), CWR(TD-RH), CWR(ASD-LH), and CWR(ASD-RH), together with a matched morphological random-connectivity control, CWR-Random(35×35). This is an ablation of the connectivity representation used to instantiate CWR; the VLM baselines are not evaluated on morphological connectomes at any point. All morphological-connectome results are reported in Table 5 alongside the functional-connectome CWR results and the independent VLM baselines, and are marked with bold square brackets to distinguish them from results already reported in the main paper.

Table 5: Complete comparison of CWR (functional and morphological connectomes), matched random-connectivity controls, and five vision-language model baselines, across all seven RAVEN tasks. Functional-connectome CWR rows (CWR-TD, CWR-ASD, CWR-Random(111×111)) and the Qwen2-VL-2B/Phi-3.5 baselines reproduce results already reported in the main paper. Newly added values — InternVL2-2B, Qwen2.5-VL-3B, PaliGemma2-3B, and the morphological-connectome CWR variants and their random control — are marked **[in bold square brackets]**.

Method	Ctr	2×2G	3×3G	L-R	U-D	O-IC	O-IG
Functional-connectome CWR							
CWR-TD	76.44 ± 3.05	41.52 ± 3.93	45.32 ± 2.16	69.87 ± 8.39	75.30 ± 6.82	67.76 ± 5.61	57.50 ± 3.39
CWR-ASD	77.81 ± 2.90	42.30 ± 3.40	45.49 ± 3.52	70.18 ± 10.05	74.62 ± 6.90	67.67 ± 5.28	56.38 ± 4.15
CWR-Random(111×111)	73.98 ± 2.23	40.33 ± 2.40	41.10 ± 1.82	63.82 ± 6.41	70.73 ± 2.72	62.91 ± 2.41	48.73 ± 4.64
Independent vision-language model baselines							
Qwen2-VL-2B	67.64 ± 1.93	41.67 ± 1.05	43.36 ± 0.55	44.57 ± 1.20	44.85 ± 1.66	47.88 ± 1.04	39.86 ± 1.80
[InternVL2-2B]	[72.03 ± 7.67]	[50.46 ± 1.27]	[52.64 ± 1.39]	[58.10 ± 1.90]	[71.93 ± 1.60]	[66.67 ± 1.01]	[62.50 ± 1.09]
[Qwen2.5-VL-3B]	[86.14 ± 1.10]	[57.10 ± 1.60]	[47.47 ± 1.96]	[89.72 ± 3.94]	[45.01 ± 3.42]	[82.35 ± 1.49]	[75.98 ± 2.16]
Phi-3.5	90.63 ± 0.83	56.23 ± 1.33	60.02 ± 1.02	87.22 ± 3.18	89.73 ± 0.72	76.73 ± 1.90	76.92 ± 2.88
[PaliGemma2-3B]	[98.63 ± 0.35]	[77.55 ± 1.08]	[81.25 ± 0.39]	[97.59 ± 0.33]	[98.26 ± 0.14]	[97.22 ± 0.33]	[94.04 ± 0.29]
[Morphological-connectome CWR]							
[CWR(TD-LH)]	[64.50 ± 6.97]	[31.73 ± 2.09]	[34.27 ± 1.49]	[48.95 ± 3.92]	[40.37 ± 4.79]	[50.54 ± 6.83]	[65.22 ± 6.34]
[CWR(TD-RH)]	[68.10 ± 4.01]	[31.10 ± 3.22]	[35.06 ± 1.64]	[49.77 ± 4.34]	[41.40 ± 3.24]	[52.17 ± 7.11]	[68.27 ± 7.09]
[CWR(ASD-LH)]	[66.66 ± 4.85]	[31.49 ± 2.51]	[35.14 ± 2.19]	[48.89 ± 3.73]	[40.92 ± 4.54]	[50.96 ± 6.86]	[66.52 ± 5.47]
[CWR(ASD-RH)]	[67.52 ± 4.64]	[31.75 ± 2.80]	[35.28 ± 1.92]	[49.79 ± 3.60]	[41.00 ± 3.51]	[53.08 ± 9.00]	[67.92 ± 6.90]
[CWR-Random(35×35)**]**	[57.16 ± 11.83]	[29.74 ± 1.85]	[33.96 ± 0.47]	[47.65 ± 4.61]	[40.74 ± 4.44]	[49.46 ± 5.68]	[61.35 ± 5.48]

Interpretation of the morphological-connectome results. As expected for a connectivity representation not directly optimized for functional integration, the morphological-connectome CWR variants generally achieve lower absolute accuracy than the strongest functional-connectome CWR results (CWR-TD, CWR-ASD). Nonetheless, the four morphological CWR variants; CWR(TD-LH), CWR(TD-RH), CWR(ASD-LH), and CWR(ASD-RH), outperform the matched morphological random-connectivity control, CWR-Random(35×35), in 27 of 28 task-level comparisons; the single exception is within one standard deviation of both conditions. This pattern mirrors the functional-connectome result reported in the main paper (Table 1), where biological connectivity outperforms matched random connectivity of the same dimensionality. Taken together, this provides additional support for *Hypothesis 3* : the computational advantage conferred by biological wiring is not tied exclusively to one particular connectivity representation, and persists, albeit at reduced absolute magnitude, when CWR is instantiated with morphology-derived rather than functional connectivity. Layer-wise CKA/RSA alignment visualizations for the three additional VLMs and for the morphological-connectome analysis are provided in Appendix H.1.

Framing the comparison with vision-language models. Several of the independent VLM baselines, most notably PaliGemma2-3B and Phi-3.5-Vision, achieve higher absolute accuracy than CWR on most tasks (Table 5). This is expected: these models bring billions of learned parameters, large-scale visual pretraining, and task-specific fine-tuning, whereas CWR's recurrent operators are entirely fixed by empirical connectivity and receive no gradient whatsoever (Appendix B); only a lightweight readout is trained. Rather than treating the VLM comparison as a contest to be won on

raw accuracy alone, we read it as evidence of a distinct and complementary property: CWR reaches competitive, non-trivial performance across two independent connectivity representations (functional and morphological) using a training-free, gradient-free reservoir and a fraction of the compute of the VLM baselines (8–10 minutes on a consumer-grade GPU vs. 5–10 hours of fine-tuning on server-class hardware; Appendix D). The central conclusion supported by Table 5 is therefore twofold: CWR's advantage over matched random connectivity is robust across connectivity representations, and this robustness is achieved with dramatically lower parametric and training requirements than the large-scale VLMs it is compared against.

F LEAK-RATE SENSITIVITY ACROSS VLM ARCHITECTURES

We evaluate the reservoir's sensitivity to the leak rate α by sweeping $\alpha \in \{0.1, 0.2, \ldots, 1.0\}$ (5 seeds per value) and recomputing task accuracy together with CKA and RSA at $\mathbf{H}_3$ against all five VLM architectures listed in Appendix E.

Averaged over the seven RAVEN tasks, accuracy decreases monotonically as α increases, from 62.33% at $\alpha = 0.1$ to 55.61% at $\alpha = 1.0$; every individual task decreases with α, with Center and U-D showing the largest absolute change (Figure 7).

CKA at $\mathbf{H}_3$, averaged over the seven tasks, changes little across the sweep for every VLM, with a relative change between $+1.9\%$ and $+6.5\%$ from $\alpha = 0.1$ to $\alpha = 1.0$, and the ranking of tasks by CKA is largely preserved across the sweep (Spearman rank correlation between the $\alpha = 0.1$ and $\alpha = 1.0$ conditions ranges from 0.61 to 0.96 across the five VLMs; Table 6). RSA at $\mathbf{H}_3$ decreases with α for all five VLMs, with slopes between -0.002 and -0.006 per unit α; for Qwen2-VL-2B, RSA changes sign over the sweep, from positive at $\alpha = 0.1$ to slightly negative at $\alpha = 1.0$. This pattern is consistent with a dissociation in which CKA is comparatively insensitive to leak rate whereas RSA is more sensitive to temporal integration, and it holds across all five architectures tested.

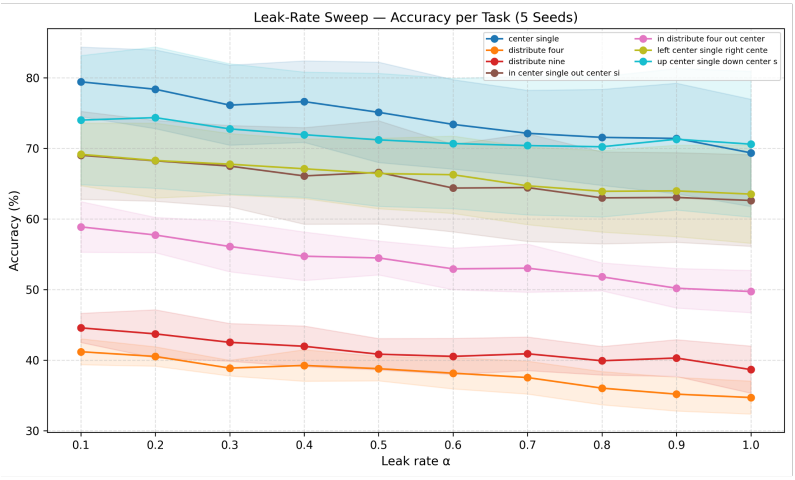


Figure 7: Task accuracy as a function of leak rate α (5 seeds per value). Accuracy decreases monotonically with α for all seven RAVEN tasks.

G NKP-RANK SENSITIVITY

We evaluate the effect of the NKP decomposition rank on representational alignment by factorizing the Center-task connectome at rank $r \in \{1, 2, 4\}$ and recomputing CKA and RSA at $\mathbf{H}_3$ (fusion mode) against all five VLM architectures listed in Appendix E. As a direct measure of how much connectome structure each rank retains, we also compute the fraction of connectome (Frobenius) variance captured at each rank: 32.67% at rank 1 (the value used throughout the main paper), 43.31% at rank 2, and 60.19% at rank 4.

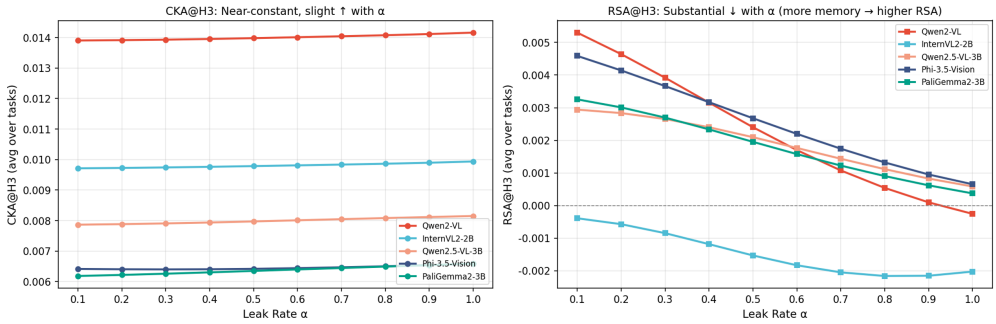


Figure 8: CKA at $\mathbf{H}_3$ (left) and RSA at $\mathbf{H}_3$ (right), averaged over the seven tasks, as a function of leak rate α, for all five VLMs.

Table 6: Leak-rate sensitivity summary per VLM. ρ: Spearman rank correlation of task-level alignment between $\alpha = 0.1$ and $\alpha = 1.0$ ($^*p < 0.05$). CKA rel. Δ: relative change in average CKA at $\mathbf{H}_3$ across the sweep. RSA abs. Δ and slope: change in average RSA at $\mathbf{H}_3$ across the sweep.

VLM	CKA ρ	CKA rel. Δ	RSA ρ	RSA abs. Δ	RSA slope/$\Delta\alpha$
Qwen2-VL-2B	0.964^*	$+1.9\%$	0.214	-0.00556	-0.00641
InternVL2-2B	0.607	$+2.3\%$	0.821^*	-0.00164	-0.00214
Qwen2.5-VL-3B	0.857^*	$+3.6\%$	0.857^*	-0.00236	-0.00280
Phi-3.5-Vision	0.750	$+2.6\%$	0.750	-0.00394	-0.00450
PaliGemma2-3B	0.750	$+6.5\%$	0.714	-0.00288	-0.00336

CKA at $\mathbf{H}_3$ increases monotonically with rank for all five VLMs, with relative increases from $r = 1$ to $r = 4$ ranging from 3.5% to 11.0%; the ordering of VLMs by CKA magnitude, highest for Qwen2-VL-2B and lowest for Phi-3.5-Vision and PaliGemma2-3B, is preserved at every rank (Figure 9). RSA at $\mathbf{H}_3$ increases with rank for Qwen2-VL-2B and PaliGemma2-3B, rising from 0.0144 to 0.0212 and from 0.0123 to 0.0149 respectively, but remains close to zero across the tested ranks for InternVL2-2B, Phi-3.5-Vision, and Qwen2.5-VL-3B on this task. The increase in CKA with rank, consistent across all five architectures, supports the interpretation of rank-1 NKP as a conservative estimate of the alignment recoverable at higher rank; the RSA result indicates that this effect on relational alignment is not uniform across VLM architectures within the range of ranks tested.

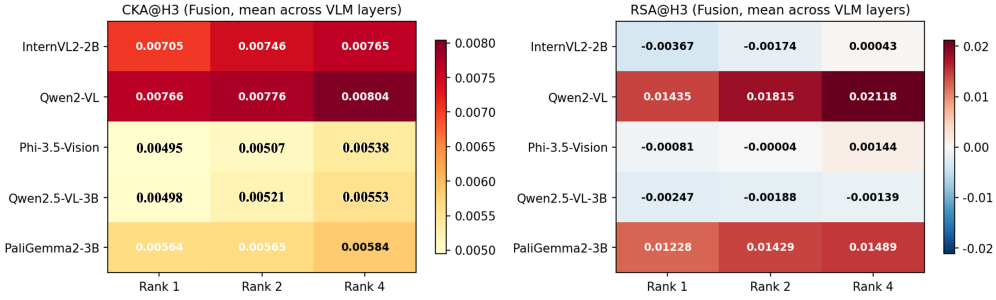


Figure 9: CKA at $\mathbf{H}_3$ (left) and RSA at $\mathbf{H}_3$ (right), fusion mode, Center task, as a function of NKP rank $r \in \{1, 2, 4\}$, for all five VLMs.

Task accuracy is comparably insensitive to NKP rank (Table 7): the seven-task average varies by less than 2 percentage points across $r \in \{1, 2, 4\}$ (rank 1: 61.54%; rank 2: 63.01%, the highest of the three; rank 4: 62.01%), indicating that, unlike alignment, task accuracy does not benefit monotonically from a higher-fidelity NKP approximation.

Table 7: Task accuracy (%, mean±SD, 5 seeds) as a function of NKP rank.

Rank	Ctr	2×2G	3×3G	L-R	U-D	O-IC	O-IG	Avg
1	78.44 ± 5.16	40.04 ± 1.60	43.95 ± 2.48	68.12 ± 5.65	73.63 ± 9.65	68.85 ± 5.78	57.73 ± 1.78	61.54
2	79.46 ± 2.02	42.34 ± 1.63	45.18 ± 2.60	71.47 ± 3.01	76.06 ± 2.84	69.27 ± 2.78	57.28 ± 1.11	63.01
4	79.08 ± 1.53	41.60 ± 1.41	45.08 ± 2.90	70.35 ± 3.10	75.00 ± 2.24	66.79 ± 1.52	56.21 ± 3.64	62.01

H Full per-task alignment results

Figures 10–12 show per-task alignment results for the remaining RAVEN configurations, computed using the TD functional connectome under Fusion and individual routes (Row 3, Column 3, Diagonal 1), for both Qwen and Phi models. For each task, CKA and RSA heatmaps (top) report alignment between reservoir steps $\{\mathbf{H}_1, \mathbf{H}_2, \mathbf{H}_3\}$ and VLM layers, while the corresponding curves (bottom) summarize layer-wise trends.

Qwen. Across tasks, CKA exhibits a consistent increase with depth, with alignment peaking in later layers. This pattern is stable across Fusion and individual routes, indicating robust feature-level alignment. In contrast, RSA shows clear route-dependent structure: Fusion yields a smoother transition toward positive alignment at deeper layers, whereas individual routes introduce intermediate anti-correlated regions—most prominently for Row 3—followed by partial recovery in later layers. This highlights that relational alignment is sensitive to the route decomposition despite stable CKA trends.

Phi. CKA also increases with depth but with sharper transitions across reservoir steps, reflecting stronger step-specific effects than in Qwen. RSA patterns are more segmented: early and intermediate layers often exhibit weak or negative alignment, followed by localized positive regions that vary across routes. Fusion produces smoother, more coherent transitions, while individual routes (notably Column 3 and Diagonal 1) lead to more fragmented alignment structure.

Summary. Across all tasks, alignment consistently strengthens toward deeper VLM layers. Fusion stabilizes alignment patterns, whereas individual routes introduce structured variability. Compared to Qwen's smooth and gradual alignment, Phi shows sharper transitions and more heterogeneous alignment regimes.

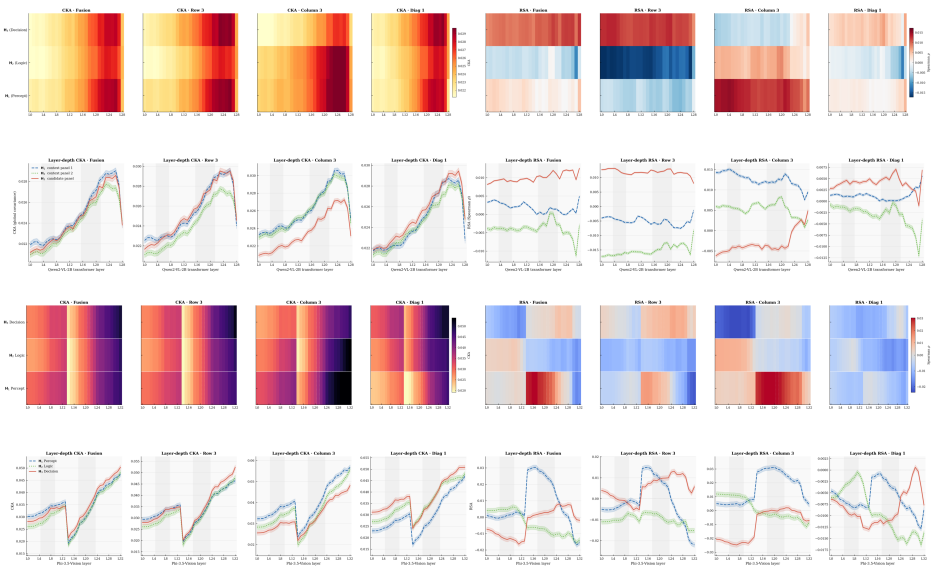
H.1 Additional alignment visualizations: new VLMs and morphological connectome

Complementing the aggregate accuracy comparison in Appendix E, this subsection reports the full per-task alignment visualizations underlying the additional VLM and morphological-connectome analyses. Figures 13–15 show CKA and RSA heatmaps and layer-depth profiles between CWR-TD reservoir steps (H_1–H_3, fusion mode) and each of the three additional VLMs — InternVL2-2B, Qwen2.5-VL-3B, and PaliGemma2-3B — across all seven RAVEN tasks, following the same protocol and layout conventions as the original-model figures in Appendix H. Figure 16 reports the corresponding alignment between the morphological-connectome CWR variants (CWR(TD-LH), CWR(TD-RH)) and InternVL2-2B, used here as a representative alignment target for this connectivity-representation ablation.

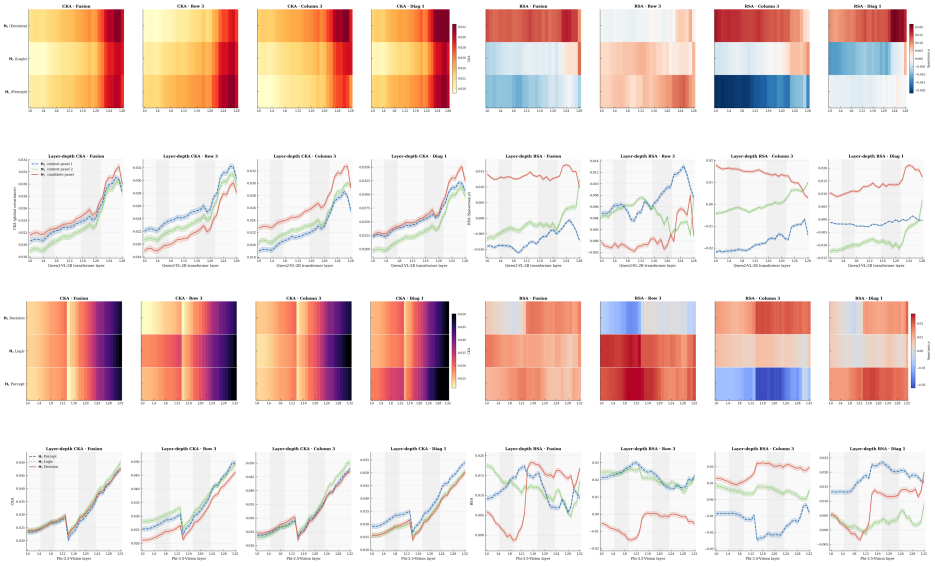
I Correctness-conditioned alignment: remaining tasks

Figures 17–18 present correctness-conditioned CKA and RSA heatmaps for the six remaining tasks, complementing the Center task analysis in Section 4.1. Because context panels are shared across all candidates, $\mathbf{H}_1$ and $\mathbf{H}_2$ are numerically identical between correct and error trials by construction; any alignment divergence is therefore necessarily localized at $\mathbf{H}_3$, providing a clean, unconfounded test of candidate-integration selectivity across tasks of varying structural complexity.

The correctness-conditioned dissociation reported for the Center task replicates consistently across all remaining tasks. Under correct trials, $\mathbf{H}_3$ systematically attains the highest representational alignment with VLM deep layers, as evidenced by both CKA covariance correspondence and the positive RSA band localized at the terminal-integration step. Under error trials, this ordering collapses: dominant



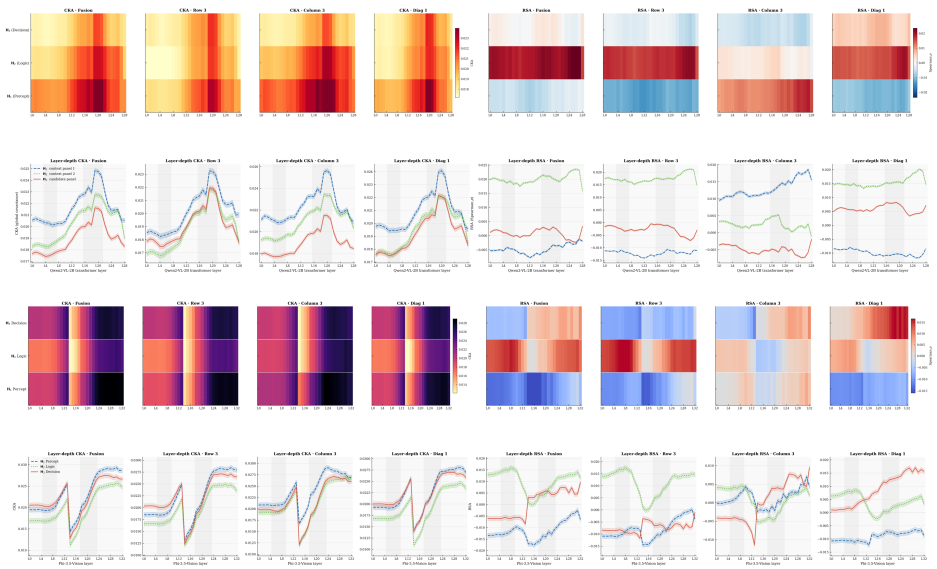
2×2GRID



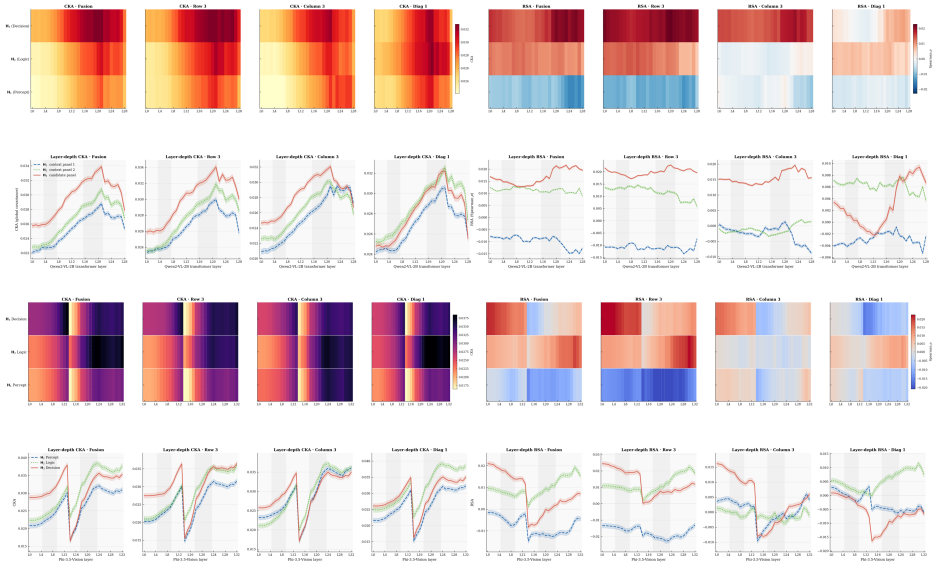
3×3GRID

Figure 10: Alignment between CWR states and VLM representations on two Raven reasoning tasks. For each task, the first two rows show Qwen2-VL-2B and the last two rows show Phi-3.5-Vision. Within each model, the first row presents layer-wise heatmaps under fusion mode across three reasoning routes, with the first four panels reporting CKA and the next four RSA; the second row shows the corresponding layer-depth profiles across transformer layers. The top pair corresponds to **2×2Grid** and the bottom pair to **3×3Grid**.

alignment migrates toward earlier CWR states, with $\mathbf{H}_1$ rising in CKA and $\mathbf{H}_2$ assuming the most positively aligned role in RSA, indicating that the reservoir fails to consolidate relational structure at the candidate-integration stage and instead strands it at an intermediate or early-integration abstraction level. The precise layer depth at which the correct-vs.-error divergence emerges varies with task complexity, simpler tasks exhibit earlier and broader separation, while structurally more demanding tasks concentrate the dissociation in deeper layers, yet the directional signature remains uniform across all conditions. Critically, this pattern replicates across both Qwen2-VL-2B and Phi-3.5-Vision, confirming that the correctness-conditioned geometry is a property of the reservoir dynamics rather



O-IC (OUT-IN CENTER)



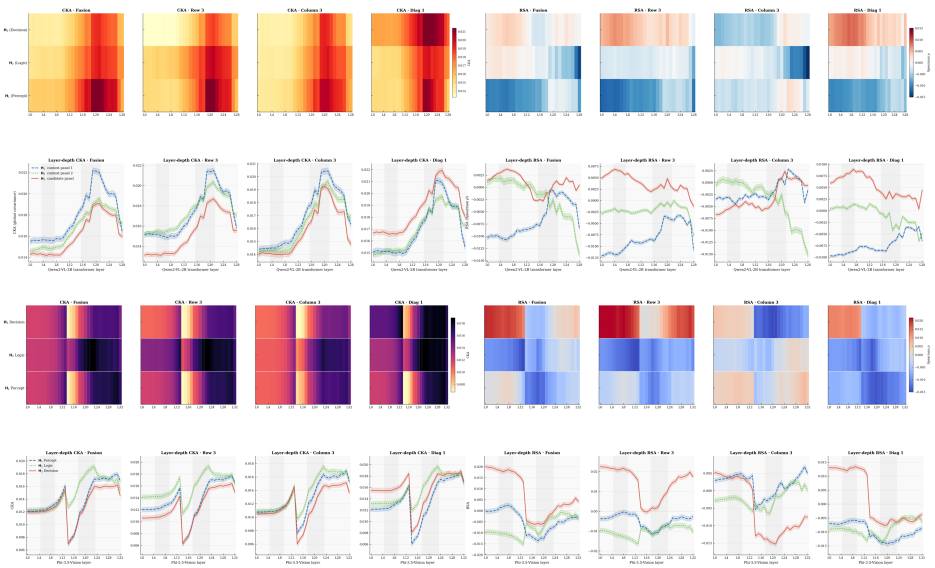
O-IG (OUT-IN GRID)

Figure 11: Alignment between CWR states and VLM representations on two Raven reasoning tasks. For each task, the first two rows show Qwen2-VL-2B and the last two rows show Phi-3.5-Vision. Within each model, the first row presents layer-wise heatmaps under fusion mode across three reasoning routes, with the first four panels reporting CKA and the next four RSA; the second row shows the corresponding layer-depth profiles across transformer layers. The top pair corresponds to **O-IC** and the bottom pair to **O-IG**.

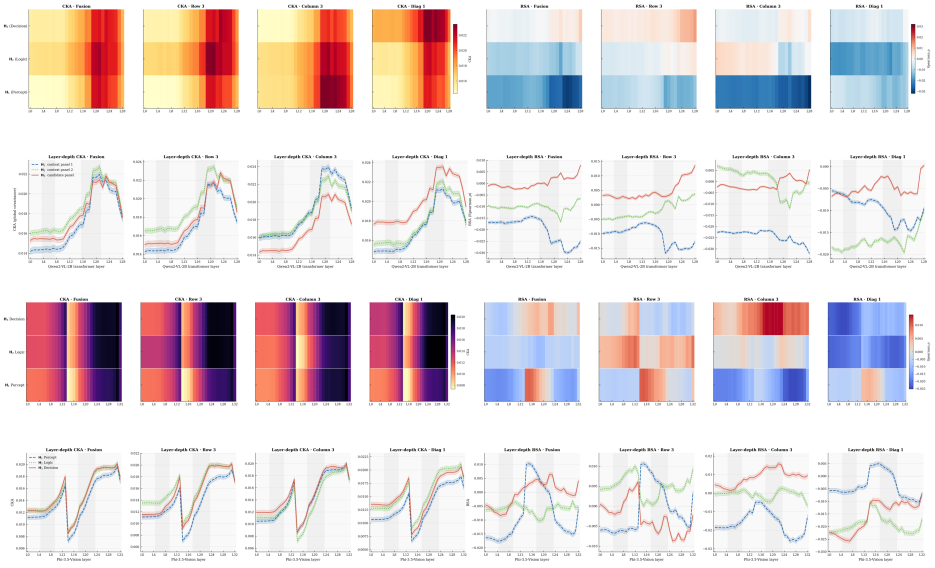
than of any specific VLM architecture. Taken together, these results support *Hypothesis 2* and establish the selectivity of $\mathbf{H}_3$ alignment for correct reasoning as a robust, task-general phenomenon.

J ABLATION STUDY: REMAINING TASKS

Figures 19–21 present the full ablation comparison across the six remaining tasks, contrasting biological connectome connectivity against random wiring and temporal integration removal (No-Leak, $\alpha = 1.0$). Together, these figures establish that the two orthogonal failure modes identified for



L-R (LEFT-RIGHT)



U-D (UP-DOWN)

Figure 12: Alignment between CWR states and VLM representations on two Raven reasoning tasks. For each task, the first two rows show Qwen2-VL-2B and the last two rows show Phi-3.5-Vision. Within each model, the first row presents layer-wise heatmaps under fusion mode across three reasoning routes, with the first four panels reporting CKA and the next four RSA; the second row shows the corresponding layer-depth profiles across transformer layers. The top pair corresponds to **L-R** and the bottom pair to **U-D**.

the Center task in the main paper are not task-specific artifacts but constitute a general property of the reservoir's representational mechanism.

Biological connectivity. Across all tasks, the connectome condition consistently produces the step-ordered representational hierarchy documented in the main paper: CKA profiles show $\mathbf{H}_3$ reaching the highest covariance alignment in deep layers, and RSA heatmaps exhibit a clean gradient from negative ($\mathbf{H}_1$) through near-zero ($\mathbf{H}_2$) to positive ($\mathbf{H}_3$), faithfully reflecting the early-to-terminal integration sequence. Layer-depth profiles confirm that this hierarchy sharpens progressively with network depth, concentrating the strongest alignment signal precisely at the candidate-integration stage.

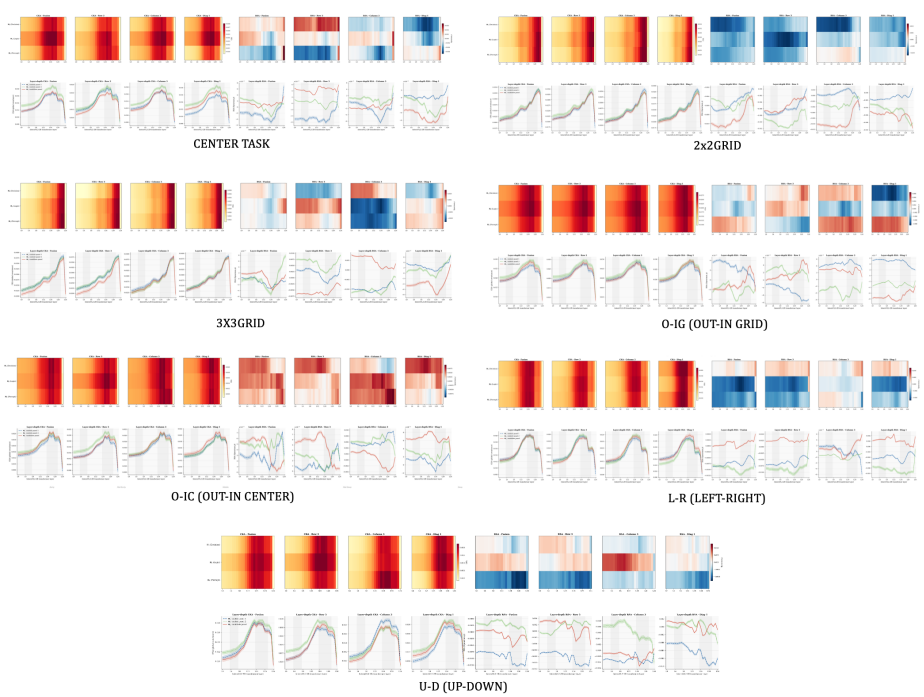


Figure 13: CKA and RSA alignment between CWR-TD reservoir steps (H_1–H_3) and InternVL2-2B layers, fusion mode, across all seven RAVEN tasks. Layout follows the per-task alignment figures in Appendix H.

Random connectivity. Replacing biological wiring with random connectivity preserves a residual CKA signal in late layers but systematically collapses the RSA geometry: the ordered gradient dissolves across all tasks into an undifferentiated pattern in which component distinctiveness is lost. This dissociation, CKA amplitude partially retained while RSA structure is abolished, confirms that biological topology contributes specifically to the *relational geometry* of representational alignment rather than to raw covariance magnitude, and that this contribution is task-general.

Temporal integration removal. Disabling cross-step accumulation ($\alpha = 1.0$) induces a qualitatively distinct failure mode that likewise replicates uniformly across tasks. CKA values inflate broadly and indiscriminately across all three CWR components, while RSA collapses into an undifferentiated negative pattern, with $\mathbf{H}_3$ losing the positive alignment that characterises correct terminal-integration encoding under the full model. Without temporal memory, the reservoir cannot accumulate the relational structure necessary to differentiate candidate integration from earlier processing steps, flattening the representational hierarchy regardless of task structure.

Summary. The consistent replication of both failure modes across tasks of varying structural complexity confirms that biological connectivity and temporal integration are jointly necessary and functionally dissociable contributors to the reservoir's alignment geometry. Neither condition alone suffices: random wiring degrades relational structure while preserving magnitude, and removing temporal integration degrades step-wise differentiation while inflating global covariance. These results provide further evidence for *Hypothesis 3* and extend its scope beyond the Center task to the full benchmark.

J.1 Model-side ablation: what kind of learned representation drives alignment?

The reservoir-side ablations in Section 4.1 establish that biological connectivity and temporal integration are jointly necessary on the *reservoir* side. A complementary question concerns the *model* side: is the observed alignment specific to representations acquired through relational learning, or would it emerge with any sufficiently deep transformer-like architecture? This question is motivated by prior findings that untrained language models with random weights already achieve substantial brain predictability (Schrimpf et al., 2021), suggesting that architectural depth alone contributes

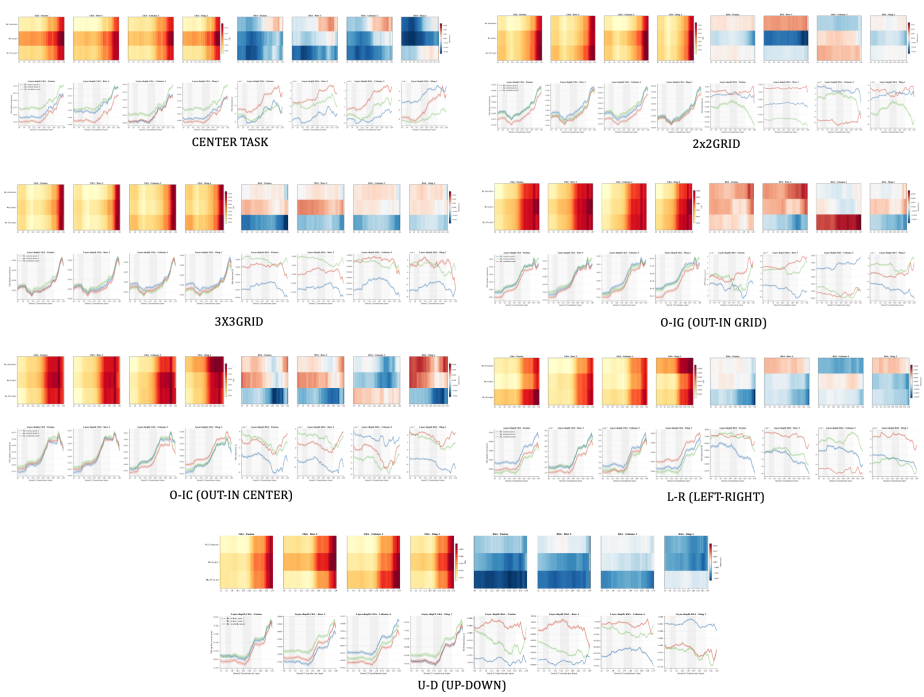


Figure 14: CKA and RSA alignment between CWR-TD reservoir steps (H_1–H_3) and Qwen2.5-VL-3B layers, fusion mode, across all seven RAVEN tasks.

to representational structure. We address this through two model-side ablations forming a double dissociation with the reservoir-side results.

Experimental setup. The CWR-TD reservoir ($\alpha = 0.2$) is held fixed throughout; only the reference model varies across three conditions: (1) **Qwen2-VL-2B fine-tuned** (main result, Section 4.1), (2) **Null-Qwen**: Qwen2-VL-2B with randomly initialized weights (identical architecture, no pretraining or fine-tuning), and (3) **ViT-B/16**: a Vision Transformer trained on image classification only (Dosovitskiy et al., 2021) (13 layers, no language or relational pretraining). Results across all seven RAVEN tasks are shown in Figure 22 (subfigures 22a–22g); findings are consistent throughout.

Null-Qwen: architectural depth alone is insufficient. Across all seven tasks (Figure 22), replacing fine-tuned Qwen weights with random initializations of identical architecture produces a flat, unstructured CKA response: the deep-layer gradient concentrated at $\mathbf{H}_3$ that characterizes the main result (Figure 2) is uniformly absent across Center (22a), 2×2Grid (22b), 3×3Grid (22c), L-R (22f), U-D (22g), O-IC (22d), and O-IG (22e) tasks, and no step-ordered hierarchy emerges at any depth. RSA likewise collapses to an undifferentiated pattern in which the ordered negative-to-positive gradient across reservoir stages is lost, with no consistent positive signature at $\mathbf{H}_3$. This outcome is task-general and diagnostic: the alignment documented in Section 4.1 is not a consequence of transformer architecture per se — of attention mechanisms, residual connections, or layer normalization — but depends specifically on *representations shaped by relational learning*. This contrasts with the language domain, where architectural depth alone yields surprisingly high brain predictability (Schrimpf et al., 2021); in the relational visual reasoning domain, architecture absent learned weights contributes nothing to the structured alignment signal.

ViT-B/16: perceptual learning yields diffuse, unordered alignment. Substituting ViT-B/16 produces a qualitatively distinct failure mode, visible consistently across all tasks in Figure 22. CKA heatmaps show a diffuse, moderate-magnitude response in which all three reservoir stages reach broadly comparable alignment values, without the $\mathbf{H}_3$-dominant hierarchy of the fine-tuned model. RSA profiles are similarly unordered: the clean negative-to-positive gradient across $\mathbf{H}_1$, $\mathbf{H}_2$, and $\mathbf{H}_3$ is absent, replaced by mixed or inverted signatures carrying no consistent correspondence to the reasoning sequence. This pattern holds uniformly from the simplest layout (22a) to the most spatially

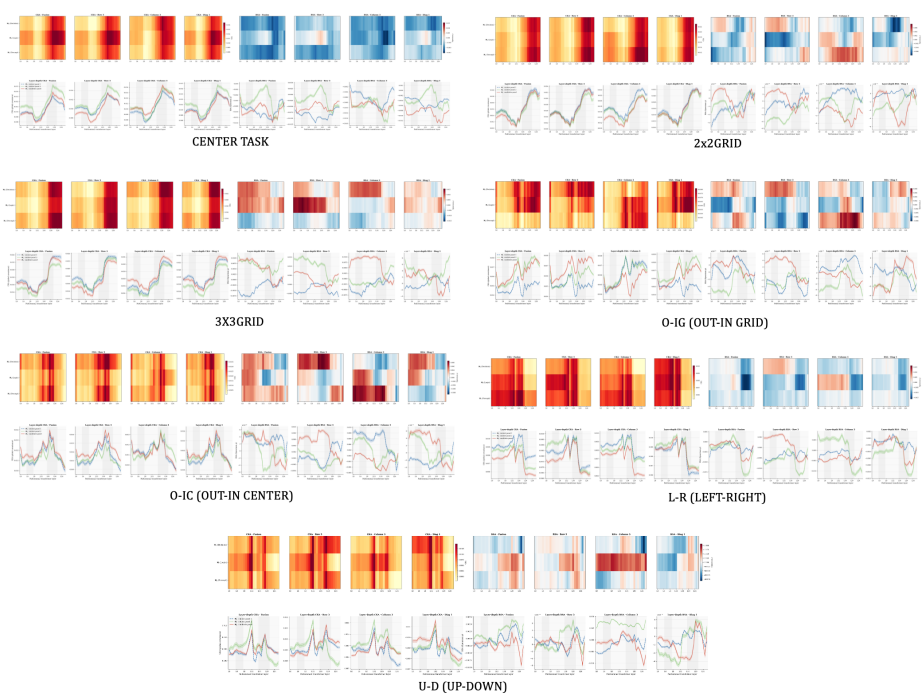


Figure 15: CKA and RSA alignment between CWR-TD reservoir steps (H_1–H_3) and PaliGemma2-3B layers, fusion mode, across all seven RAVEN tasks.

complex tasks (22e, 22f, 22g), indicating that low-level visual features captured by ViT partially overlap with early reservoir encoding ($\mathbf{H}_1$ (early integration)) but the absence of abstract relational learning prevents the emergence of terminal-integration alignment at $\mathbf{H}_3$. Recovering the step-ordered representational hierarchy requires not merely visual processing but representations oriented toward relational and compositional reasoning — a capacity that image-classification pretraining does not provide (Barrett et al., 2018).

A complete double dissociation. Jointly, the reservoir-side and model-side ablations yield a clean double dissociation that holds across the full task battery. On the reservoir side, removing biological topology or temporal integration degrades RSA structure while leaving CKA partially intact, implicating biological wiring specifically in relational rather than geometric alignment. On the model side, removing learned weights abolishes alignment entirely, while substituting perceptual-only learning recovers diffuse covariance without step-ordered relational structure. The intersection of these two dissociations identifies the necessary and sufficient conditions for the observed alignment: biologically-wired reservoir dynamics resonating with representations acquired through abstract relational learning. This result reduces the plausibility of two parsimonious alternative explanations, namely that alignment is an artifact of transformer depth or that it reduces to shared low-level visual processing, and is consistent with *Hypothesis 3* .

K Statistical validation

K.1 Pairwise comparisons: CWR-TD vs. vectorized RC baselines

Table 8 reports raw accuracy for CWR against three vectorized reservoir-computing baselines (ESN, LSM, and a fractal multi-timescale reservoir; Appendix B.1) under an identical hyperparameter and readout architecture, with reservoir connectivity as the only variable. CWR outperforms all three baselines on every task, for both the TD and ASD connectomes.

Table 9 reports pairwise two-sided t-tests and Cohen's d effect sizes comparing CWR-TD against each vectorized RC baseline (ESN-TD, LSM-TD, Fractal-TD) across all seven tasks ($n = 5$ seeds). For the 3×3Grid task, the CWR-TD comparisons with ESN-TD ($p = 0.0101$) and Fractal-TD

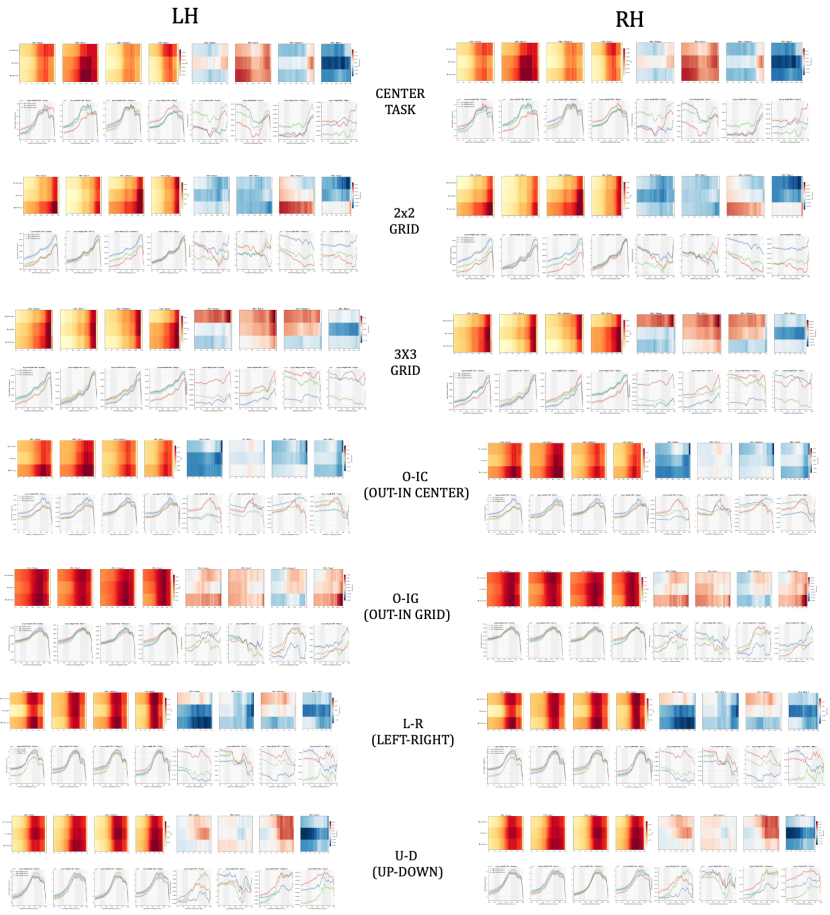


Figure 16: CKA and RSA alignment between the morphological-connectome CWR variants (CWR(TD-LH), CWR(TD-RH), CWR(ASD-LH), CWR(ASD-RH)) and InternVL2-2B layers, complementing the accuracy results reported in Table 5 (Appendix E).

Table 8: Visual reasoning accuracy (%) across RAVEN tasks: CWR vs. vectorized reservoir-computing (RC) baselines. TD and ASD are group-level functional connectomes; Random = matched spectral radius and density.

Model	CBT	Ctr	2×2G	3×3G	L-R	U-D	O-IC	O-IG
ESN (Jaeger, 2001)	TD	65.33 ± 1.71	38.98 ± 0.56	41.22 ± 1.16	54.53 ± 0.90	54.57 ± 2.21	56.64 ± 1.68	45.83 ± 1.47
	ASD	64.13 ± 1.85	38.22 ± 0.79	40.65 ± 1.39	54.62 ± 1.62	54.49 ± 1.59	57.05 ± 1.45	46.17 ± 0.74
	Random	65.62 ± 1.42	38.77 ± 0.38	41.13 ± 0.87	53.89 ± 1.10	53.42 ± 1.39	56.90 ± 1.06	46.12 ± 0.90
LSM (Maass et al., 2002)	TD	39.63 ± 1.54	34.10 ± 1.08	35.53 ± 1.31	45.31 ± 1.72	45.40 ± 2.05	47.15 ± 1.47	39.00 ± 0.76
	ASD	39.78 ± 0.92	34.45 ± 1.49	34.83 ± 1.45	46.15 ± 1.16	45.23 ± 2.01	47.75 ± 0.85	39.17 ± 0.72
	Random	39.72 ± 2.77	34.46 ± 0.55	35.07 ± 2.31	45.51 ± 1.97	45.24 ± 1.93	47.33 ± 1.50	38.73 ± 1.56
Fractal	TD	65.82 ± 1.68	39.47 ± 1.06	42.24 ± 1.27	54.41 ± 1.36	53.90 ± 1.14	56.80 ± 1.48	46.70 ± 0.84
	ASD	65.35 ± 2.08	39.44 ± 0.87	41.83 ± 0.94	54.85 ± 1.55	54.10 ± 1.43	57.12 ± 0.80	47.34 ± 0.97
	Random	65.69 ± 2.87	39.81 ± 0.82	41.74 ± 0.55	54.91 ± 2.12	53.39 ± 1.20	57.29 ± 0.82	46.43 ± 1.53
CWR	TD	76.44 ± 3.05	41.52 ± 3.93	45.32 ± 2.16	69.87 ± 8.39	75.30 ± 6.82	67.76 ± 5.61	57.50 ± 3.39
	ASD	77.81 ± 2.90	42.30 ± 3.40	45.49 ± 3.52	70.18 ± 10.05	74.62 ± 6.90	67.67 ± 5.28	56.38 ± 4.15
	Random	73.98 ± 2.23	40.33 ± 2.40	41.10 ± 1.82	63.82 ± 6.41	70.73 ± 2.72	62.91 ± 2.41	48.73 ± 4.64

($p = 0.00393$) are statistically significant. The nonsignificant comparisons include 2×2Grid vs. ESN-TD ($p = 0.237$), 2×2Grid vs. Fractal-TD ($p = 0.343$), L-R vs. ESN-TD ($p = 0.0664$), and 3×3Grid vs. LSM-TD ($p = 0.0541$). Given $n = 5$, these tests have limited statistical power, so effect sizes and confidence intervals should be considered alongside the p-values. Across all tasks and baseline comparisons, effect sizes range from moderate to very large, with the strongest advantages observed against LSM-TD (e.g., Center: $d = 13.63$; O-IG: $d = 6.74$; see Table 9 for the complete set).

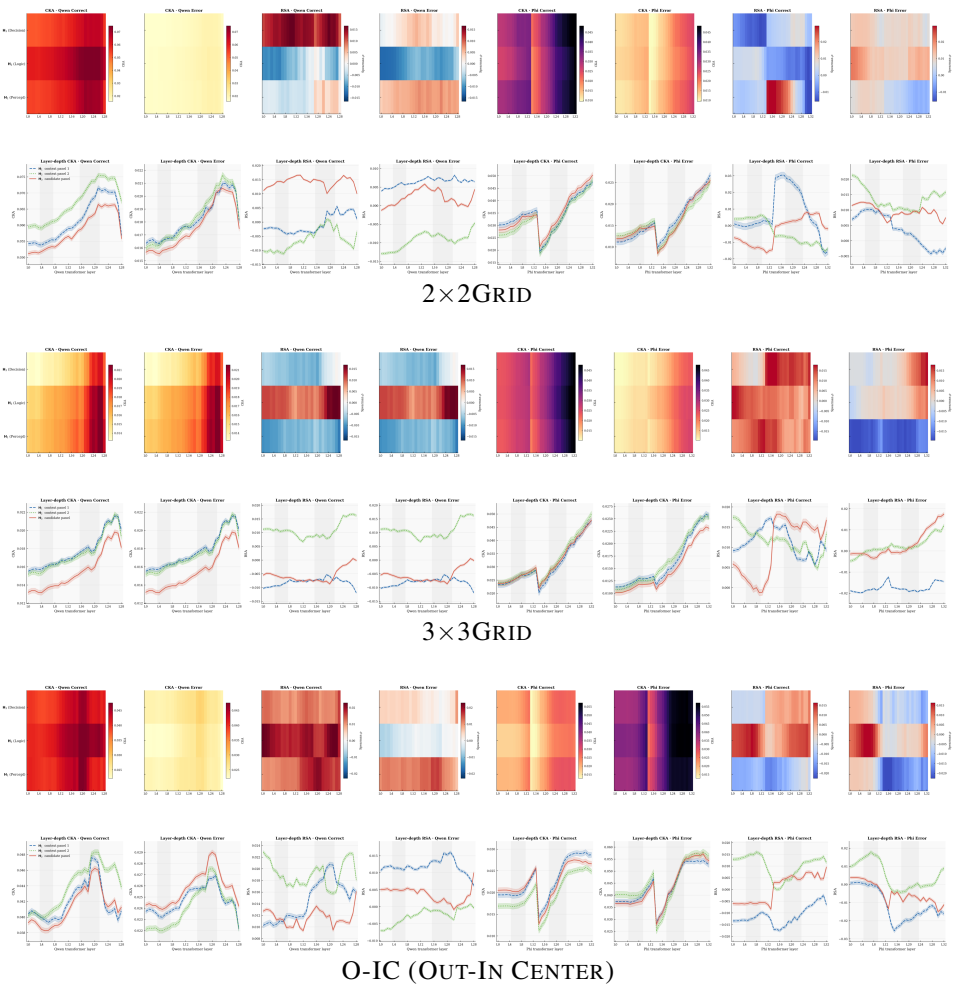
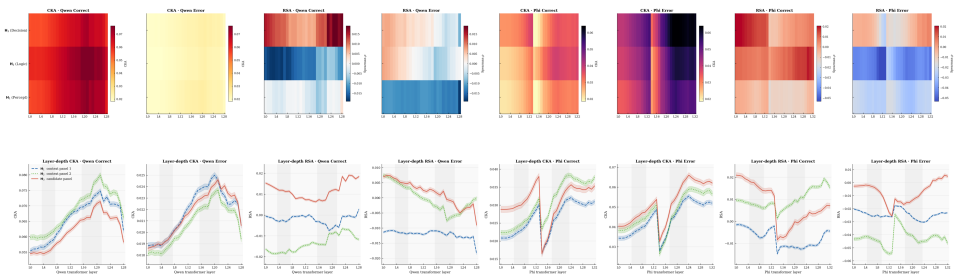


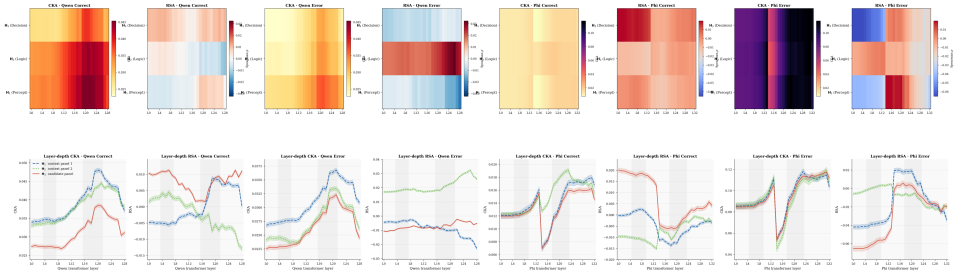
Figure 17: Correctness-conditioned alignment between CWR reservoir steps (H_1–H_3) and VLM representations on three grid-based Raven tasks. For each panel: **Top section** shows CKA and RSA heatmaps comparing CWR dynamics against Qwen2-VL-2B and Phi-3.5-Vision layer-wise features for correct vs. error candidates; **Bottom section** displays corresponding layer-depth alignment profiles. All analyses use fusion mode + three reasoning routes (Row 3, Column 3, Diag 1). 2×2GRID and 3×3GRID test scaling of distributed reasoning; O-IC evaluates generalization from localized input to localized output. Stronger alignment indicates tighter representational correspondence during successful reasoning.

Table 9: Pairwise t-tests and Cohen's d: CWR-TD vs. vectorised RC baselines ($n=5$ seeds). Non-significant p-values at limited sample size should be interpreted alongside effect sizes, which range from moderate to very large.

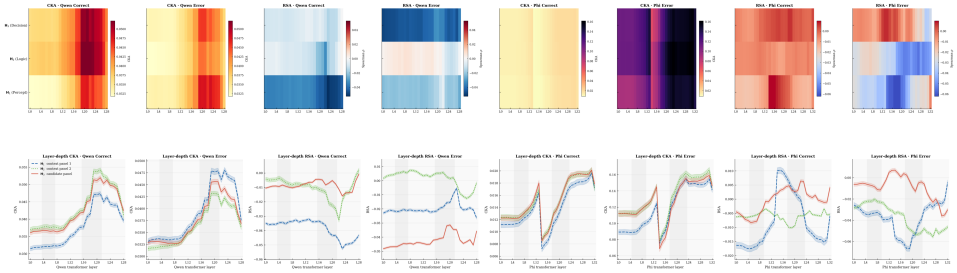
	vs. ESN-TD		vs. LSM-TD		vs. Fractal-TD	
Task	p	d	p	d	p	d
Center	2.19×10^{-4}	4.02	2.27×10^{-5}	13.63	2.89×10^{-4}	3.86
2×2Grid	2.37×10^{-1}	0.81	6.59×10^{-3}	2.30	3.43×10^{-1}	0.64
3×3Grid	1.01×10^{-2}	2.12	5.41×10^{-2}	4.91	3.93×10^{-3}	1.56
O-IC	5.25×10^{-3}	2.40	1.01×10^{-4}	4.49	5.41×10^{-3}	2.39
O-IG	2.27×10^{-4}	4.00	5.27×10^{-6}	6.74	2.63×10^{-4}	3.91
L-R	6.64×10^{-2}	2.30	4.37×10^{-4}	3.63	6.22×10^{-3}	2.30
U-D	4.15×10^{-4}	3.66	3.08×10^{-5}	5.31	2.63×10^{-3}	3.91



O-IG (OUT-IN GRID)



L-R (LEFT-RIGHT)



U-D (UP-DOWN)

Figure 18: Correctness-conditioned alignment between CWR reservoir steps (H_1–H_3) and VLM representations on three generalization/spatial Raven tasks. For each panel: **Top section** shows CKA and RSA heatmaps comparing CWR dynamics against Qwen2-VL-2B and Phi-3.5-Vision layer-wise features for correct vs. error candidates; **Bottom section** displays corresponding layer-depth alignment profiles. All analyses use fusion mode + three reasoning routes (Row 3, Column 3, Diag 1). O-IG tests input distribution generalization; L-R and U-D evaluate horizontal vs. vertical spatial reasoning. Alignment strength reflects how well reservoir dynamics mirror VLM representations during structured, correctness-dependent inference.

K.2 ASD VS. TD GROUP-LEVEL COMPARISON

Table 10 reports Cohen's d effect sizes comparing CWR instantiated with ASD- and TD-derived connectome templates against CWR with random connectivity. Across all tasks, both groups yield positive effect sizes, confirming that biologically grounded connectivity consistently confers a reasoning advantage over random reservoirs, irrespective of the clinical group from which the template is derived.

The magnitude of this advantage varies systematically with task structure. The largest effects are observed on tasks requiring globally structured relational reasoning — 3×3Grid (ASD: $d = 1.40$, TD: $d = 1.89$) and O-IG (ASD: $d = 1.55$, TD: $d = 1.93$) — while tasks with simpler spatial configurations yield more moderate gains, and the 2×2Grid task produces the smallest effects overall, particularly for TD ($d = 0.33$). This task-dependent gradient is consistent with the interpretation that biological connectivity introduces a geometry-sensitive inductive bias whose benefit scales with the relational demands of the task, rather than providing a uniform performance increment.

Differences between ASD and TD templates are secondary to their shared advantage over random connectivity and vary in direction across tasks: ASD yields a larger effect on the Center task ($d = 1.33$ vs. 0.82), while TD achieves stronger gains on globally structured transformations such as O-IG and U-D. Crucially, both groups exhibit consistent, non-negligible improvements across all tasks, with effect sizes largely overlapping in the medium-to-very-large range. These results support the conclusion that the performance advantage of CWR over random connectivity is attributable to the shared properties of biological connectivity structure, rather than to group-specific features of either template. This comparison is included strictly as a robustness check on whether the reservoir's advantage over random connectivity generalizes across cohorts; it is not a claim about cognitive ability or diagnostic status associated with either cohort.

Table 10: Cohen's d: CWR-TD and CWR-ASD vs. CWR-random, with effect size classifications. We use the conventional Cohen's d thresholds 0.2 (small), 0.5 (moderate), 0.8 (large), and 1.2 (very large). Task-specificity of the advantage is consistent with a geometry-sensitive inductive bias introduced by biological connectivity (see §4.2).

Task	ASD d	ASD Class	TD d	TD Class
Center	1.33	Very Large	0.82	Large
2×2Grid	0.60	Medium	0.33	Small
3×3Grid	1.40	Very Large	1.89	Very Large
O-IC	1.04	Large	1.00	Large
O-IG	1.55	Very Large	1.93	Very Large
L-R	0.67	Medium	0.72	Medium
U-D	0.66	Medium	0.79	Medium

K.3 Robustness to option-order shuffling

Table 11 reports option-order robustness for CWR and both VLM baselines across all tasks. For every model and every task, the shuffling p-value exceeds 0.05, indicating no statistically detectable sensitivity to option order under this test. CWR's Relative Performance Fluctuation (RPF) remains at or below 3.07% across all seven tasks, with the largest value on 3×3Grid ($\mathrm{RPF} = 3.07\%$, $p = 0.37$) and every other task below 1.3%; none approaches significance, consistent with option orderings introducing only surface-level variability without affecting the underlying performance distribution. Both Qwen2-VL-2B and Phi-3.5-Vision exhibit uniformly low RPF values ($\leq 3.22\%$) across all tasks, indicating that option-order invariance is a property of the evaluation setting rather than of any specific model. Taken together, these results confirm that the performance rankings reported in the main paper are stable to permutations of the candidate set and are not attributable to positional bias.

K.4 Resampling methodology

Since the biological and random connectomes are each a single fixed matrix reused across all five seeds, the reservoir component of the alignment does not vary across seeds; treating stimulus-level observations as independent replicates would inflate the effective sample size. For orderings across reservoir stages and across NKP ranks, we therefore use a permutation test that shuffles stage labels within each problem instance, plus bootstrap confidence bands built on the same resampling unit. For quantities with genuine seed-level variation (correctness-conditioned alignment at $\mathbf{H}_3$, task accuracy, both downstream of the trained readout), we instead use paired permutation tests across the five seeds; with five seeds the exact sign-flip permutation space is $2^5 = 32$, so we report effect sizes and confidence intervals alongside these tests rather than relying on p-values alone.

L Individual connectome analysis

The main paper employs group-level connectomes derived separately for ASD and TD cohorts. A natural question is whether the reported alignment and performance results are specific to the group-level average, or whether they generalize to individual connectome topology, including subjects whose structural organization deviates substantially from the group mean. This section addresses

Table 11: Option-shuffling robustness (mean over 5 seeds). RPF = Relative Performance Fluctuation (%). No statistically detectable sensitivity to option order was observed under this test.

Model	Task	Orig	Shuff	Diff	RPF (%)	p
CWR	Center	76.44	76.91	+0.47	0.61	0.8638
	2×2Grid	41.52	42.04	+0.52	1.25	0.8572
	3×3Grid	45.32	43.93	−1.39	3.07	0.3665
	O-IC	67.76	68.03	+0.27	0.40	0.9456
	O-IG	57.50	57.96	+0.46	0.80	0.8470
	L-R	69.87	69.94	+0.07	0.10	0.9913
	U-D	75.30	75.57	+0.27	0.36	0.9585
Qwen2-VL	Center	66.74	65.91	−0.83	1.24	0.5785
	2×2Grid	42.48	41.68	−0.80	1.88	0.2568
	3×3Grid	43.43	42.53	−0.90	2.07	0.1852
	O-IC	47.68	46.14	−1.54	3.22	0.1304
	O-IG	39.87	40.67	+0.80	2.00	0.4225
	L-R	44.80	44.31	−0.49	1.09	0.5110
	U-D	44.76	43.97	−0.79	1.76	0.5824
Phi-3.5	Center	90.63	91.60	+0.97	1.07	0.3206
	2×2Grid	56.23	56.47	+0.24	0.43	0.7454
	3×3Grid	60.02	60.46	+0.44	0.73	0.4846
	O-IC	76.73	78.36	+1.63	2.12	0.2142
	O-IG	76.92	76.34	−0.58	0.75	0.7896
	L-R	87.22	86.15	−1.07	1.23	0.5657
	U-D	89.73	88.96	−0.77	0.86	0.4555

that question through two complementary analyses: a performance robustness check using the most topologically atypical individuals, and a representational alignment analysis examining whether the depth-ordered CKA and RSA signatures replicate at the individual level.

Furthest-subject robustness. To probe robustness under maximal topological deviation, we identify the three ASD and three TD subjects with the largest Frobenius distance from their respective group-level connectomes — hereafter ASD-Far1–3 and TD-Far1–3 — and instantiate a separate CWR reservoir for each. Table 12 reports reasoning accuracy across all seven tasks for these six individuals. Despite representing the extreme tail of the within-group structural distribution, all six subjects sustain consistent accuracy levels that closely track the group-level results reported in the main paper, with no subject falling below 39% on the most demanding tasks and performance on simpler tasks (Center, U-D, L-R) remaining above 63% throughout. The absence of systematic degradation at maximum topological distance confirms that CWR reasoning performance is robust to individual structural variability and does not depend on proximity to the group-level connectome.

Table 12: CWR reasoning accuracy (%, mean $\pm$ std over 5 seeds) for the six individuals most topologically distant from their respective group-level connectomes. Results confirm performance robustness under maximal within-group structural deviation.

Task	ASD-Far1	ASD-Far2	ASD-Far3	TD-Far1	TD-Far2	TD-Far3
Center	73.73 ± 1.64	75.11 ± 1.65	75.94 ± 1.40	74.18 ± 2.58	73.06 ± 2.06	74.89 ± 1.63
2×2Grid	39.18 ± 1.49	40.79 ± 2.03	43.41 ± 1.30	41.07 ± 2.63	40.91 ± 2.61	40.44 ± 2.18
3×3Grid	42.05 ± 1.55	41.45 ± 1.40	42.21 ± 1.63	41.40 ± 1.58	41.73 ± 2.18	40.54 ± 1.64
O-IC	64.11 ± 4.01	67.89 ± 3.00	69.97 ± 4.73	65.34 ± 1.65	66.39 ± 5.81	63.78 ± 5.54
O-IG	47.70 ± 3.88	49.62 ± 4.15	50.83 ± 3.09	47.18 ± 1.35	48.72 ± 6.64	47.71 ± 3.67
L-R	63.18 ± 4.58	67.74 ± 1.88	72.55 ± 2.46	64.39 ± 4.46	66.86 ± 6.12	63.87 ± 2.66
U-D	72.39 ± 4.63	76.52 ± 3.04	80.23 ± 3.83	74.88 ± 3.88	73.30 ± 4.66	73.63 ± 3.52

Individual-level representational alignment. Figures 23–24 present CKA and RSA heatmaps and layer-depth profiles for all six furthest subjects against Qwen2-VL-2B representations, shown here for a representative task (U-D); results across remaining tasks are consistent. Despite substantial topological distance from the group-level connectome and meaningful variation in absolute alignment magnitudes across individuals, the core representational signature of the main paper replicates at the individual level in both cohorts. Under CKA, $\mathbf{H}_3$ (terminal integration) attains the highest covariance alignment in deep layers across subjects, with the step-ordered hierarchy preserved in the layer-depth profiles. Under RSA, the ordered gradient — with $\mathbf{H}_3$ carrying the dominant positive relational alignment in deep layers — is consistently present, though the depth at which it crystallizes and its absolute magnitude vary across individuals as a function of their specific connectivity topology. Notably, this pattern holds irrespective of diagnostic group: ASD-Far and TD-Far subjects exhibit qualitatively identical alignment signatures, differing in degree rather than in kind. This replication at the individual level, under conditions of maximal topological atypicality, supports the interpretation that the depth-ordered CWR–VLM correspondence is a stable property of cortical organization across individuals and clinical profiles, rather than an artifact of group-level averaging or a marker of any specific population.

M Additional connectivity and preprocessing controls

This appendix consolidates selected control analyses that target the central threat to Hypothesis 3: that the CWR–VLM alignment reflects generic graph structure, a preprocessing artifact, or the trained readout, rather than the empirical connectome per se. Each control below isolates one specific alternative explanation.

Structure-matched graph nulls. We evaluated four nulls beyond the uniform-random baseline used in the main paper: degree-preserving Maslov–Sneppen rewiring, ROI-permuted connectivity (which preserves the empirical connectome's values and spectrum but destroys their regional arrangement), a stochastic block model (SBM), and a Watts–Strogatz small-world graph (Table 13). On *accuracy*: CWR vs. uniform random, $+4$–5 points; CWR vs. Maslov–Sneppen, comparable (within ~ 1 point); CWR vs. ROI-permuted, 61.96% vs. 57.70%; CWR vs. SBM/Watts–Strogatz, comparable (within ~ 1 point). On *alignment*: CKA is higher for the empirical connectome in 63/63 matched comparisons against both Maslov–Sneppen and ROI-permuted connectivity. Moreover, BioFC/null mean CKA ratios at the terminal-integration stage $\mathbf{H}_3$ were approximately 1.31–$1.32\times$ (Maslov–Sneppen), 1.28–$1.31\times$ (SBM), 1.19–$1.21\times$ (Watts–Strogatz), and 2.10–$2.23\times$ (region-shuffled); the region-shuffled comparison was significant in all 7 tasks for each of the three VLMs, and the advantage over the other structured nulls was observed across most tasks. A separate spectrum-matched synthetic (non-biological) operator was exceeded by the empirical connectome at $\mathbf{H}_3$ (macro-mean CKA ratio 1.05–$1.10\times$ across two VLMs, empirical higher in 5/7 and 6/7 tasks respectively; Table 13), while region-shuffled connectivity, which preserves that same spectrum, lost most of the alignment advantage. Together these indicate that spectral structure alone is neither irrelevant nor sufficient, and that the empirical connectome's regional arrangement contributes an advantage beyond what a matched spectrum captures on its own.

Connectivity-control comparison. We compare independently defined connectivity controls rather than treating them as a progressive degradation sequence. Accuracy remains comparable across several structured controls, while empirical FC shows distinct representational alignment characteristics. For the positions-permuted condition, the Stage-3 ($\mathbf{H}_3$) CKA reduction relative to the empirical connectome is approximately 2.10–$2.23\times$. These results reduce the plausibility that the observed alignment is explained by preserving only a generic spectrum or overall weight statistics.

Sign processing. We tested how much of the effect depends on the exact biological sign pattern of the connectivity matrix. Macro-averaged accuracy: all-positive (absolute value only) 37.26%; original biological sign placement 50.44%; absolute value + random sign (the pipeline used throughout the main paper) 61.96%; sign-shuffled with the same positive/negative proportion preserved 64.70%. Removing all negative weights causes a large performance collapse, showing that mixed-sign recurrence is important; however, the exact biological placement of signs is *not* necessary for strong performance, since both engineered sign conditions outperform the original biological sign placement. Notably, the absolute-value-plus-random-sign condition (used throughout this paper)

has higher CKA despite lower accuracy than the sign-shuffled condition, the condition with the best decoding accuracy is not the condition with the strongest representational correspondence, so we treat accuracy and alignment as separate outcomes throughout.

Interpolation control. Replacing the resize-by-interpolation step with zero-padding, which cannot mix unrelated ROIs, while keeping the rest of the pipeline and all five alignment-analysis seeds fixed, leaves depth-wise alignment profiles nearly unchanged across two VLMs, seven tasks, and three reservoir stages (median correlation 0.998, minimum 0.949). Zero-padding preserves the qualitative cross-depth CKA profile, indicating that linear ROI mixing is not necessary for the observed depth structure; however, preprocessing affects absolute alignment magnitude and decoding accuracy. The corresponding accuracy comparison ($72.60\% \rightarrow 65.58\%$) comes from a dedicated three-seed ablation setting and should be interpreted only within that evaluation setting.

Alternative connectome construction (morphological connectome). We separately rebuilt the reservoir from cortical morphology rather than resting-state functional connectivity – a different biological data source, construction pipeline, and native resolution, requiring no resizing step. The morphological connectome outperformed its matched random control in 27 of 28 task-level comparisons (the single exception was within one standard deviation of both conditions); full accuracy figures are reported in Appendix E, Table 5. This shows the result persists when the entire connectome construction pipeline is changed, not only when individual preprocessing steps are ablated.

Readout invariance. Because the recurrent operators never receive a gradient, the reservoir states $\{\mathbf{H}_1, \mathbf{H}_2, \mathbf{H}_3\}$ analyzed throughout this paper are, in principle, independent of readout training. We verified this directly: across all seven tasks and five seeds, stored reservoir states are numerically identical before and after readout training, with a maximum difference of zero (see also Appendix B). This is the empirical basis for treating the representational-alignment results in §4.1 as independent of the readout's training objective.

CKA estimator robustness. To address the concern that our central CKA comparisons could be artifacts of the feature-to-sample ratio or of high dimensionality, we recomputed the empirical-vs.-Maslov–Sneppen CKA difference under instance subsampling (250/500/1,000 instances), under a shared 128-dimensional PCA reduction of both representations, and under a debiased CKA estimator. The sign of the empirical-connectome advantage was preserved in every case for both Qwen2-VL-2B and Phi-3.5-Vision, and the same ordering persisted after reducing both representations to 128 PCA dimensions; the BioFC-vs.-region-shuffled difference was even larger and likewise remained positive. Under the debiased estimator, the mean difference remained positive but was smaller in magnitude than under the raw estimator, and several task-level confidence intervals crossed zero.

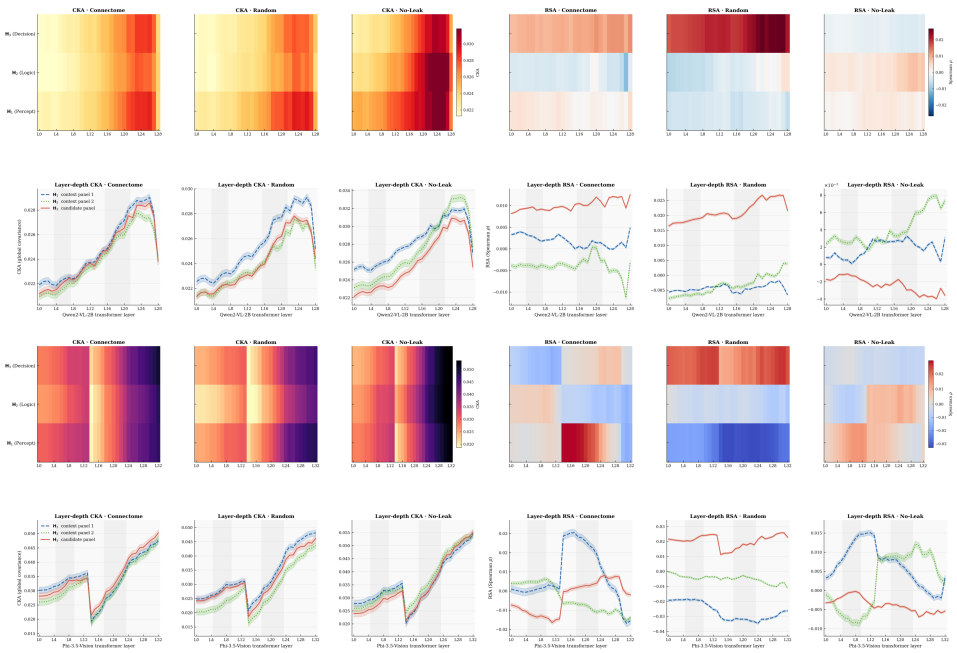
An independently constructed set of null connectivity matrices. As a further, independently run check on the specificity of the empirical connectome, we instantiated the reservoir with three additional non-biological control matrices, a within-subject regional permutation of the empirical connectome (Perm-Matrix), a weight-shuffled variant that preserves the connectome's weight distribution while randomizing which edge each weight is assigned to (WeightShuffle-Matrix), and a synthetic matrix matched to the empirical connectome's spectral profile (SpectralMatch-Matrix), and compared them against the TD-mean empirical connectome on the Center task with InternVL2-2B as the alignment target. Perm-Matrix and WeightShuffle-Matrix achieve accuracy within approximately one percentage point of the empirical connectome, but their CKA at $\mathbf{H}_3$ is low and shows no structured gradient (TD-mean CKA at $\mathbf{H}_3 \approx 0.014$ vs. Perm-Matrix ≈ 0.006, a $2.5\times$ difference; Table 13). SpectralMatch-Matrix behaves differently: it reaches CKA magnitudes comparable to the empirical connectome's own ($\mathbf{H}_3$ CKA in the 0.18–0.24 range across two VLMs), but the empirical connectome still exceeds it by 1.05–$1.10\times$, higher in 5 of 7 and 6 of 7 tasks respectively. This is consistent with, and independently corroborates, the structure-matched-null results above: accuracy is comparable across connectivity conditions while representational alignment is not.

Summary. Table 13 consolidates the controls above. Together, these selected control analyses reduce the plausibility of several specific, parsimonious alternative explanations for the CWR–VLM alignment reported in §4.1: matched density, matched degree sequence, matched community or small-world structure, matched spectral profile alone, the interpolation step, the exact sign pattern, and

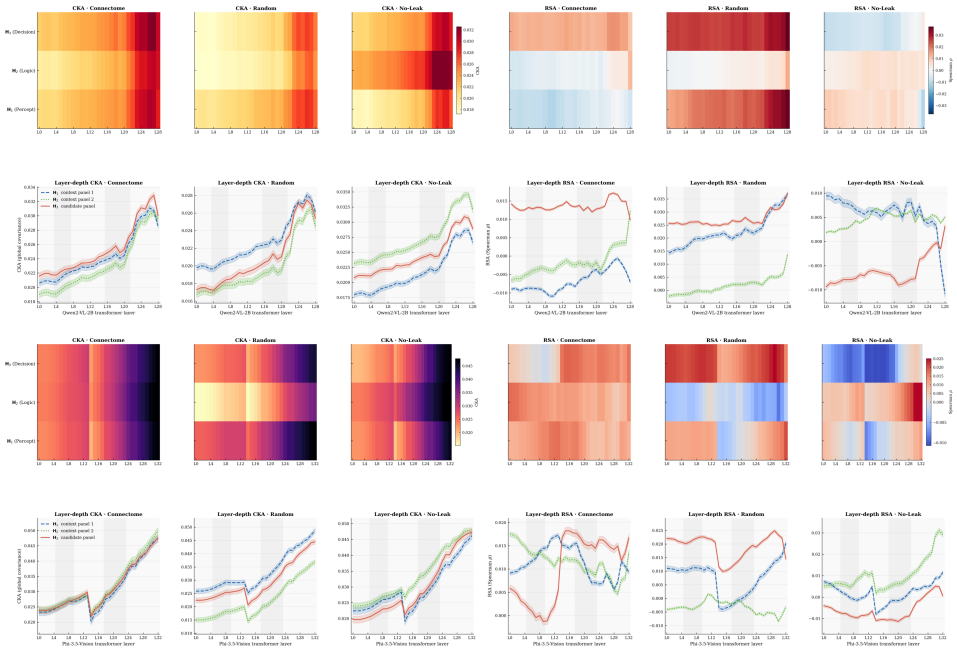
the trained readout all fail to reproduce the empirical connectome's full alignment advantage. They do not, on their own, identify which specific topological property is causally responsible; we return to this as an open question in §5. The morphological-connectome robustness check is conceptually distinct from these controls, it tests a second, independently derived *biological* connectome rather than a non-biological control matrix and is reported separately in §4.1 and Appendix E.

Table 13: Summary of connectivity-matrix and preprocessing controls (Appendix M). Accuracy is macro-averaged over the seven RAVEN tasks (%); for reference, the empirical connectome (CWR) attains 61.96%. Alignment reports raw $\mathbf{H}_3$ CKA (empirical vs. control) per VLM where available, with the empirical-to-control ratio in parentheses; a ratio > 1 favors the empirical connectome. **Bolded rows** are controls that match or exceed CWR on accuracy yet still lag on alignment, the clearest evidence that the alignment advantage does not reduce to an accuracy advantage. Reference row (CWR against itself) included for interpretability. Results are drawn from control-specific matched experiments. Interp/ZeroPad accuracy uses three seeds, whereas alignment profiles use five seeds. SpectralMatch accuracy and matched-item CKA come from separate runs. Absolute CKA values should therefore be interpreted within rows rather than across control families. Interp/ZeroPad accuracy comes from a dedicated ablation setting and should not be directly compared with the main-paper reference accuracy; comparisons should only be made within the same evaluation setting.

Alternative explanation	**Control**	**Acc. (%)**	**Alignment (CKA at $\mathbf{H}_3$, empirical vs. control, ratio)**
Reference	CWR (empirical connectome)	61.96	$1.00\times$ (self)
Density / spectral radius	Uniform random (matched)	57.37	Qwen2-VL: 0.04537 vs. 0.04202 ($1.080\times$) Phi-3.5-Vision: 0.02306 vs. 0.02106 ($1.095\times$) InternVL2-2B: 0.01906 vs. 0.01767 ($1.079\times$)
Degree sequence	Maslov–Sneppen rewiring	61.79	Qwen2-VL: 0.00655 vs. 0.00497 ($1.317\times$) Phi-3.5-Vision: 0.00499 vs. 0.00380 ($1.313\times$) InternVL2-2B: 0.00612 vs. 0.00468 ($1.306\times$)
ROI ordering / preprocessing	ROI-permuted ($\mathbf{PCP}^\top$)	57.70	Qwen2-VL: 0.00655 vs. 0.00296 ($2.211\times$) Phi-3.5-Vision: 0.00499 vs. 0.00238 ($2.099\times$) InternVL2-2B: 0.00612 vs. 0.00275 ($2.227\times$)
Community structure	**Stochastic block model**	62.80	**Qwen2-VL:** 0.00655 **vs.** 0.00501 ($1.307\times$) **Phi-3.5-Vision:** 0.00499 **vs.** 0.00388 ($1.283\times$) **InternVL2-2B:** 0.00612 **vs.** 0.00470 ($1.301\times$)
Small-world structure	**Watts–Strogatz**	62.58	**Qwen2-VL:** 0.00655 **vs.** 0.00548 ($1.197\times$) **Phi-3.5-Vision:** 0.00499 **vs.** 0.00411 ($1.214\times$) **InternVL2-2B:** 0.00612 **vs.** 0.00516 ($1.187\times$)
Regional arrangement (independent check)	Perm-Matrix	59.59	InternVL2-2B: ≈ 0.014 vs. ≈ 0.006 ($\approx 2.5\times$; Center only)
Weight placement (independent check)	WeightShuffle-Matrix	60.55	Qwen2-VL: 0.01391 vs. 0.01263 ($1.101\times$) Phi-3.5-Vision: 0.00641 vs. 0.00575 ($1.113\times$) InternVL2-2B: 0.00973 vs. 0.00849 ($1.146\times$)
Spectral profile (independent check)	SpectralMatch-Matrix	59.81	Matched test items ($N = 402$), macro-mean $\mathbf{H}_3$ fusion CKA: Qwen2-VL: 0.19318 vs. 0.18420 ($1.049\times$; empirical higher in 5/7 tasks) Phi-3.5-Vision: 0.24204 vs. 0.21929 ($1.104\times$; empirical higher in 6/7 tasks)
Low-rank / smoothing artifact	NKP rank $1\to2\to4$	61.54 / 63.01 / 62.01	Qwen2-VL: $0.01391 \to 0.01487 \to 0.01558$ Phi-3.5-Vision: $0.00641 \to 0.00693 \to 0.00718$ InternVL2-2B: $0.00973 \to 0.01023 \to 0.01053$ (increases with rank)
Interpolation artifact	Interp. $\to$ zero-padding	72.60 $\to$ 65.58 (-7.02 pp)	Qwen2-VL: $0.00830 \to 0.00678$ (-18.31% rel.) InternVL2-2B: $0.00778 \to 0.00646$ (-16.97% rel.)
Readout learning	State comparison, pre/post readout training	N/A	Reservoir states unchanged after training: $\max \lvert\Delta\mathbf{H}\rvert = 0$ across 7 tasks $\times$ 5 seeds; CKA/RSA unchanged for fixed VLM targets and estimators.

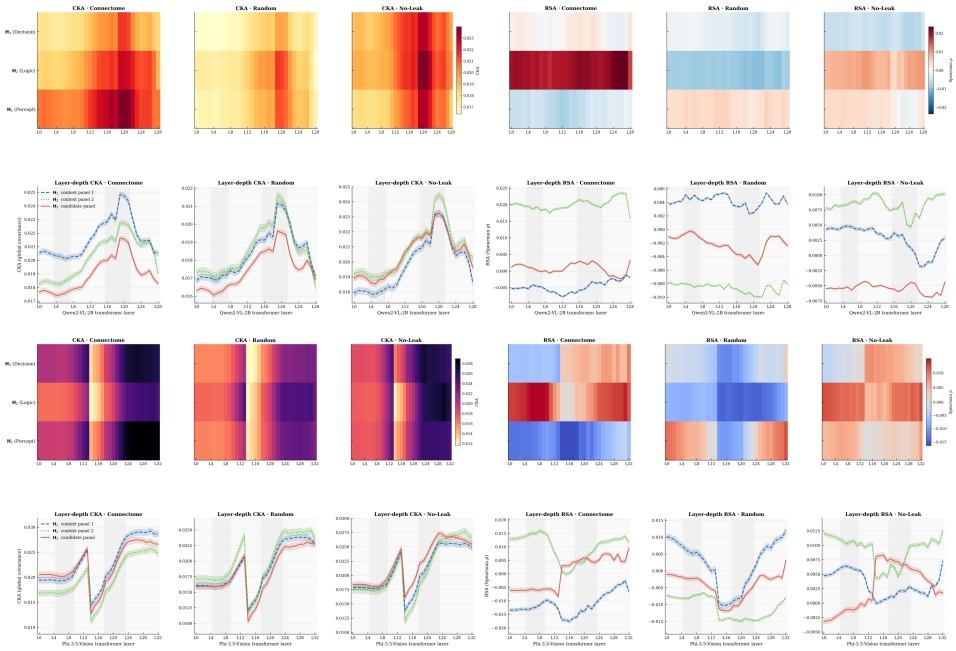


2×2GRID

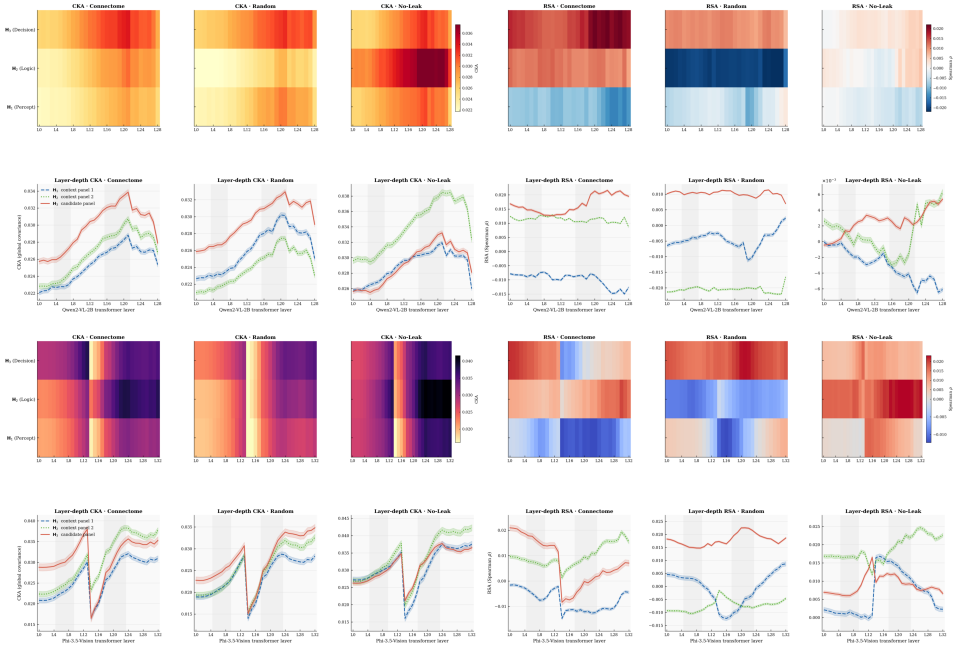


3×3GRID

Figure 19: Ablation study: Representational alignment between CWR reservoir steps and VLM layers across connectivity conditions (Connectome vs. Random vs. No-Leak) on two grid-based Raven tasks. For each task panel: **Top image** shows Qwen2-VL-2B alignment results; **Bottom image** shows Phi-3.5-Vision results. Each panel displays CKA and RSA heatmaps across transformer layers for correct vs. error candidates (top section), with corresponding layer-depth alignment profiles below (bottom section). All analyses use fusion mode + three reasoning routes (Row 3, Column 3, Diag 1). 2×2GRID and 3×3GRID evaluate how reservoir connectivity affects alignment scaling with grid complexity. Stronger alignment under Connectome configuration indicates biologically-inspired wiring better supports structured reasoning transfer.

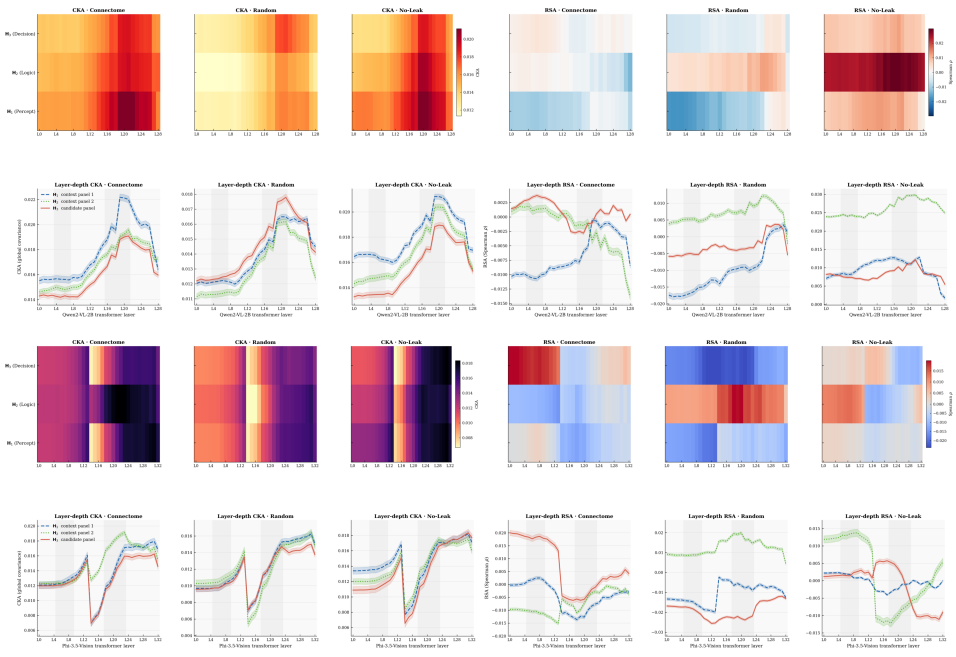


O-IC (OUT-IN CENTER)

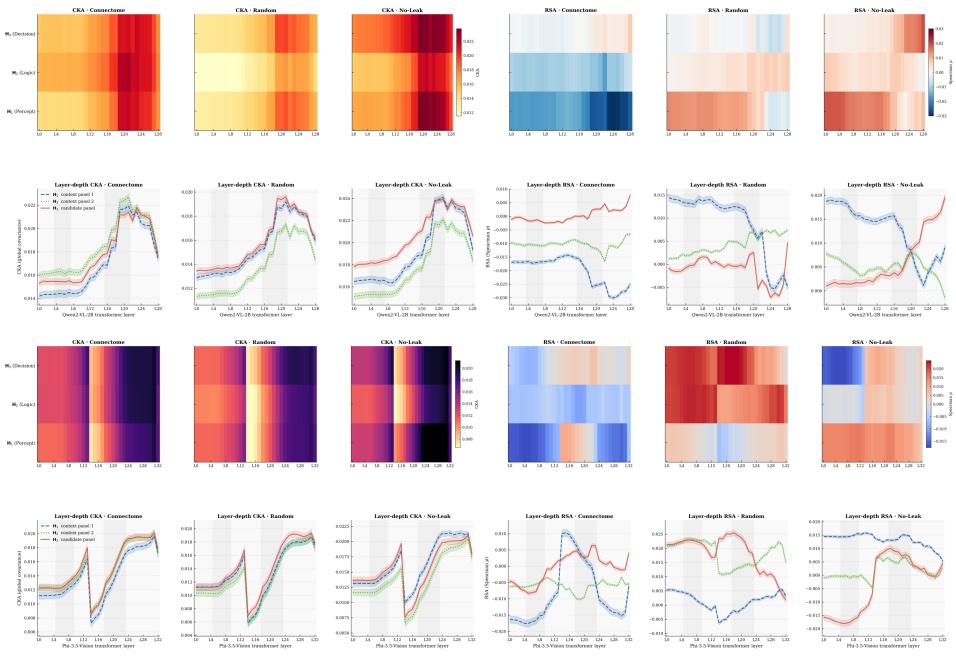


O-IG (OUT-IN GRID)

Figure 20: Ablation study: Representational alignment between CWR reservoir steps and VLM layers across connectivity conditions (Connectome vs. Random vs. No-Leak) on two generalization-focused Raven tasks. For each task panel: **Top image** shows Qwen2-VL-2B alignment results; **Bottom image** shows Phi-3.5-Vision results. Each panel displays CKA and RSA heatmaps across transformer layers for correct vs. error candidates (top section), with corresponding layer-depth alignment profiles below (bottom section). All analyses use fusion mode + three reasoning routes (Row 3, Column 3, Diag 1). O-IC tests localized input-output mapping; O-IG evaluates generalization from distributed input to localized output. Alignment differences across connectivity conditions reveal how reservoir wiring influences transfer of reasoning patterns under varying input distributions.

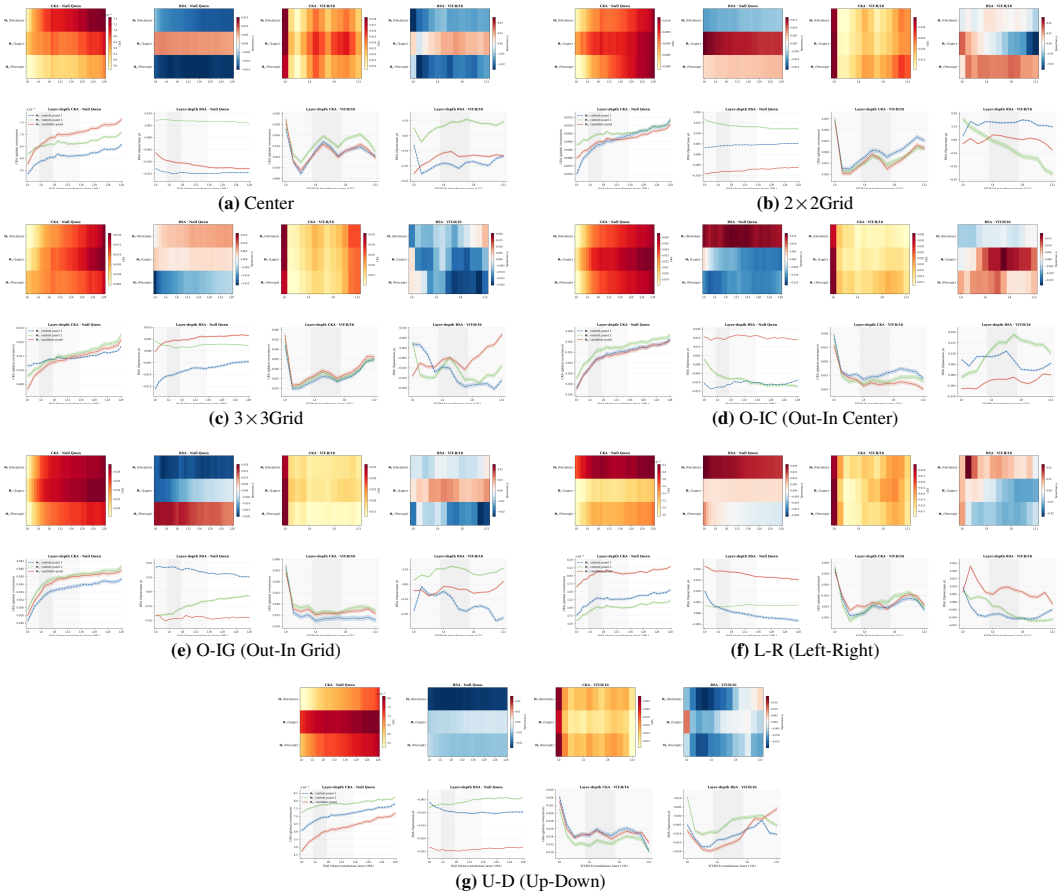


L-R (LEFT-RIGHT)



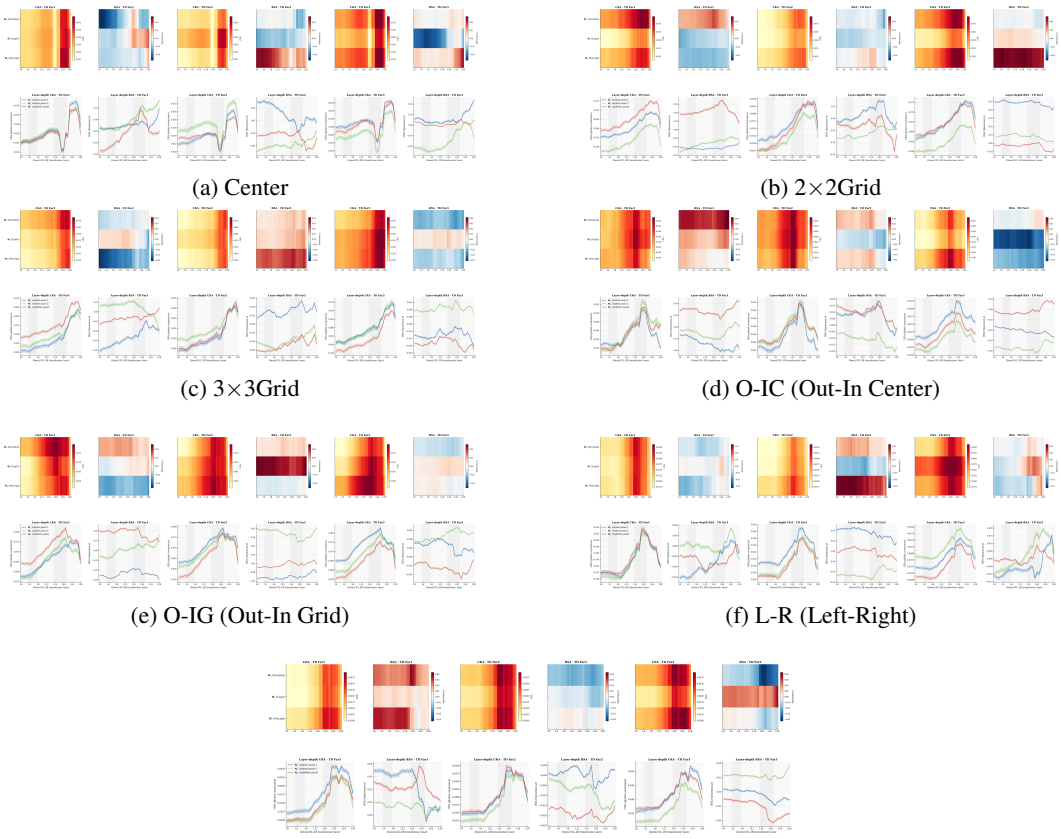
U-D (UP-DOWN)

Figure 21: Ablation study: Representational alignment between CWR reservoir steps and VLM layers across connectivity conditions (Connectome vs. Random vs. No-Leak) on two spatial-axis Raven tasks. For each task panel: **Top image** shows Qwen2-VL-2B alignment results; **Bottom image** shows Phi-3.5-Vision results. Each panel displays CKA and RSA heatmaps across transformer layers for correct vs. error candidates (top section), with corresponding layer-depth alignment profiles below (bottom section). All analyses use fusion mode + three reasoning routes (Row 3, Column 3, Diag 1). L-R evaluates horizontal-axis reasoning; U-D evaluates vertical-axis reasoning. Comparing alignment strength across connectivity conditions reveals whether biologically-inspired reservoir wiring better supports axis-specific structured inference.



(a) Center **(b)** 2×2Grid **(c)** 3×3Grid **(d)** O-IC (Out-In Center) **(e)** O-IG (Out-In Grid) **(f)** L-R (Left-Right) **(g)** U-D (Up-Down)

Figure 22: Model-side ablation: Alignment specificity comparing TD connectome against baseline architectures (Null Qwen-29L and ViT-B/16-13L) across all Raven tasks. Each panel shows CKA and RSA heatmaps (top) and layer-depth alignment profiles (bottom) for CWR reasoning steps (H_1–H_3) in TD fusion mode.



(a) Center

(b) 2×2Grid

(c) 3×3Grid

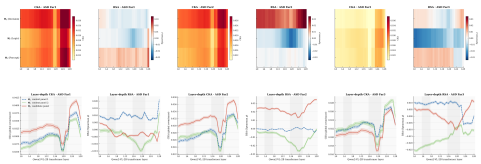
(d) O-IC (Out-In Center)

(e) O-IG (Out-In Grid)

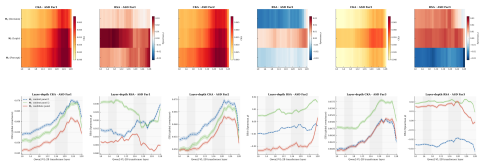
(f) L-R (Left-Right)

(g) U-D (Up-Down)

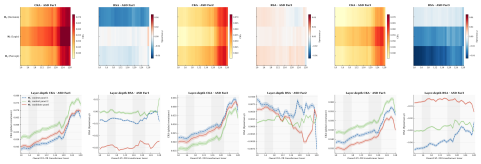
Figure 23: **Individual TD connectome alignment with Qwen2-VL-2B across all RAVEN tasks.** Each panel displays CKA/RSA heatmaps (top) and layer-depth curves (bottom) for three TD subjects (Far1, Far2, Far3). Color scales are shared within each metric (CKA/RSA) across all subjects and tasks. Region labels (Early → Deep) indicate transformer layer groupings (Qwen2-VL-2B, 29 layers). **Rows 1–2**: Tasks increasing in contextual complexity. **Rows 3–4**: Tasks with spatial manipulations of target location.



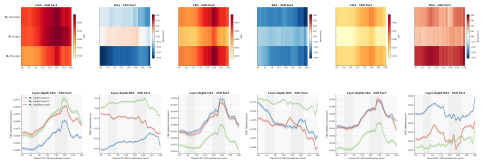
(a) Center



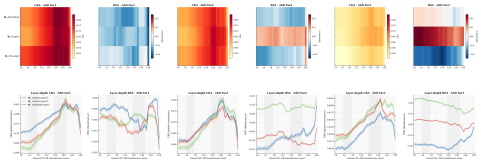
(b) 2×2Grid



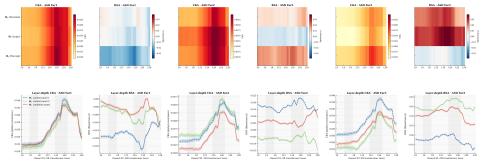
(c) 3×3Grid



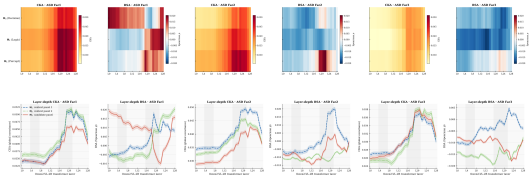
(d) O-IC (Out-In Center)



(e) O-IG (Out-In Grid)



(f) L-R (Left-Right)



(g) U-D (Up-Down)

Figure 24: **Individual ASD connectome alignment with Qwen2-VL-2B across all RAVEN tasks.** Layout, conventions, and task descriptions identical to Figure 23. Direct visual comparison with TD results reveals group-level differences in representational alignment patterns across cognitive demands. Color scales are shared within each metric (CKA/RSA) across all ASD subjects and tasks.