BRAIN-VISA: BRAIN-VISION ALIGNMENT VIA SELF-SUPERVISED BRAIN REPRESENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Human brain activity contains rich perceptual and semantic information that could complement pretrained models, yet exploiting this information is challenging. We investigate whether brain–vision alignment can improve the representations and generalization of pretrained vision models using paired image–fMRI data alone, without any behavioral supervision or category information. To enable vision models to benefit from brain–vision alignment, we leverage self-supervised learning to enhance the conceptual and semantic structure of brain representations. Specifically, we propose a two-stage **Brain**-**Vis**ion **A**lignment (Brain-VisA) framework. First, a joint-embedding predictive architecture (JEPA) learns structured brain representations through self-supervision, using repeated fMRI responses to the same image as natural positive views. Second, relational alignment jointly adapts brain and visual representations while retaining brain self-supervision. We find that the learned brain representations exhibit semantic concept structure across eight pretrained vision models from SimCLR, CLIP, and DINOv2, despite the absence of semantic supervision. Specifically, Brain-VisA improves one-shot classification accuracy for high-level semantic concepts by an average of 13.70% over the unadapted models. It further exceeds the category-informed brain–vision alignment baseline by an average of 10.14% on high-level abstract semantic classification, while requiring less training data and time. The aligned models also improve on generalization tasks including out-of-distribution recognition and the odd-one-out task. These results suggest that self-supervised learning can extract useful representational structure from limited, noisy brain signals, and that brain–vision alignment can improve the semantic representation structure and generalization of pretrained vision models without relying on any behavioral or category information.

1 INTRODUCTION

Learning visual representations from human-provided supervision has advanced substantially, from category labels and annotations to large-scale image–text pairs. Despite their different forms, these signals can all be viewed as externalized expressions of human knowledge: labels reflect how humans partition the visual world into concepts, annotations encode interpretations of visual content, and language conveys richer semantic knowledge. From this perspective, conventional supervised learning can be broadly viewed as indirect brain–vision alignment, where knowledge represented in the human brain is first externalized and then used to shape visual representations. Brain signals, however, arise directly from the internal processing of visual stimuli and may preserve information that externalized supervision cannot fully capture, particularly the latent perceptual, semantic, and relational structure through which the brain organizes visual stimuli. This motivates a more direct question: can vision models learn directly from neural responses without relying on externalized intermediates? Such direct brain–vision alignment is promising but substantially more challenging and underexplored. Compared with relatively clean, explicit, and scalable labels or text, brain signals are high-dimensional, noisy, limited in spatiotemporal resolution, and expensive to acquire. Consequently, paired image–brain datasets remain small, making it difficult to extract stable, visually relevant representations from neural measurements.

Recent studies have demonstrated the value of incorporating human representational structure into vision models. Behavioral measurements, such as human similarity judgments (Hebart et al., 2019;

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.

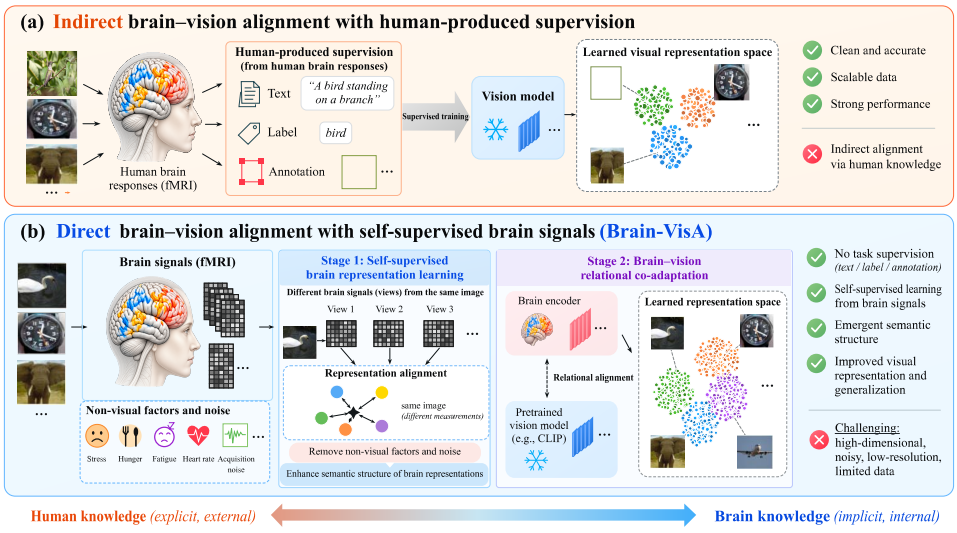

Figure 1: **Indirect and direct routes to brain–vision alignment.** Human-produced labels, text, and annotations externalize knowledge before it supervises visual learning. Neural measurements offer a more direct source of supervision, but stimulus-related information is mixed with physiological and measurement variability. Our approach learns consistency across repeated responses to the same image and uses the resulting brain representations to guide visual adaptation, without additional behavioral or category supervision during adaptation. The detailed architecture is shown in Figure 2.

Muttenthaler et al., 2025), have been used to align visual representations with human conceptual organization, improving semantic structure and generalization. Extending this direction, the mental representation-guided supervision (MRS) (Chen et al., 2026) uses representational structure derived from brain signals for brain–vision alignment and demonstrates improved visual generalization, showing that internal neural representations can benefit visual learning. However, MRS additionally introduces category information to expand the training data during alignment, making it difficult to disentangle the contribution of brain signals from that of category supervision. This raises a more fundamental question: can direct brain–vision alignment improve the semantic representations and generalization of pretrained vision models using only paired image–brain data, without behavioral supervision or category information? Answering this question would provide a more direct test of whether brain signals themselves contain representational structure that can benefit vision models.

A key challenge is learning high-quality representations from limited and noisy brain signals. If brain representations encode measurement noise, trial-specific variation, or other stimulus-irrelevant information, these factors may also be transferred to visual representations during alignment. Self-supervised learning (SSL) provides a natural solution by exploiting the intrinsic structure of brain data without additional annotations. However, the choice of objective is critical: reconstruction may preserve measurement noise alongside stimulus-related information, whereas contrastive learning with explicit negatives may separate stimuli that share meaningful semantic relationships. Effective brain self-supervision should instead suppress trial-specific variation while preserving stable stimulus-related conceptual and semantic structure.

Motivated by this goal, we introduce **Brain-Vis**ion **A**lignment (Brain-VisA), a two-stage brain-guided visual representation learning framework. In the first stage, we use a joint-embedding predictive architecture (JEPA)-based self-supervised objective and treat repeated brain responses to the same image as natural positive views to learn structured brain representations without semantic labels. Because repeated responses share the same visual stimulus while differing in measurement noise and trial-specific variability, learning consistency across them encourages the model to extract stable stimulus-related information and organize brain representations according to conceptual and semantic structure. In the second stage, we perform brain–vision relational alignment, aligning the relational structure of the learned brain representation space with that of pretrained vision models so that the structure extracted from brain signals can guide visual representation learning.

Our experiments show that self-supervised learning alone organizes brain signals into a representation space with high-level conceptual and semantic structure, and that this structure can further improve pretrained vision models through brain–vision relational alignment. Across eight pretrained vision models from SimCLR, CLIP, and DINOv2, our approach consistently improves high-level abstract semantic classification and further enhances performance on generalization tasks including odd-one-out and out-of-distribution recognition. JEPA-based brain representation learning also achieves stronger overall performance than reconstruction-based and contrastive self-supervised alternatives. Compared with a brain–vision alignment baseline assisted by category information, MRS, our method achieves higher average performance without using any behavioral supervision or category information, while requiring less training data and training time. Moreover, we observe continued improvements from brain–vision alignment as the paired image–brain data increases.

Together, these results suggest that self-supervised learning provides a powerful approach for extracting useful representational structure from limited and noisy brain signals, and that direct brain–vision alignment without behavioral supervision or category information can leverage this structure to improve the semantic representations and generalization of pretrained vision models. More broadly, our findings suggest that vision models can benefit not only from externalized forms of human knowledge, such as labels and language, but also directly from the internal representational structure through which the human brain encodes the visual world.

2 Related Work

2.1 Self-Supervised Representation Learning

Reconstruction and instance discrimination provide two major approaches to self-supervised representation learning. Autoencoders reconstruct their inputs, while denoising autoencoders recover uncorrupted targets from artificially corrupted inputs (Hinton & Salakhutdinov, 2006; Vincent et al., 2008). For fMRI, this emphasizes fidelity to a measured voxel pattern that contains both stimulus-related information and measurement noise. Instance-discriminative contrastive learning (Chen et al., 2020) instead attracts augmented views of the same instance and repels other instances, although instance-based negatives can include semantically related samples. These objectives provide useful baselines for studying which information brain self-supervision should preserve.

Negative-free methods offer alternatives through architectural asymmetry or explicit representation constraints. BYOL and SimSiam avoid negative pairs, while Barlow Twins and VICReg directly regulate feature correlations and variances (Grill et al., 2020; Chen & He, 2021; Zbontar et al., 2021; Bardes et al., 2022). These approaches motivate reducing architectural complexity while controlling representation geometry beyond avoiding constant outputs. Within this broader focus on representation-space learning, joint-embedding predictive architectures (JEPAs) learn through latent prediction rather than input reconstruction (LeCun et al., 2022; Assran et al., 2023). LeJEPA combines latent agreement with isotropic Gaussian regularization using a shared encoder and projector, without a momentum teacher or stop-gradient target branch (Balestriero & LeCun, 2025). VISReg uses sliced-Wasserstein matching to regulate distributional shape (Wu et al., 2026). Our brain SSL builds on this combination of latent agreement and explicit distributional regularization.

2.2 Neuro-AI Alignment for Improving Visual Models

Neural recordings can serve as training signals rather than only benchmarks for evaluating brain similarity (Schrimpf et al., 2021; 2020; Goldstein et al., 2022). Studies using mouse and macaque visual responses show that neural supervision can improve visual representations and robustness (Li et al., 2019; Federer et al., 2020; Safarani et al., 2021; Dapello et al., 2023). Human neural data have likewise been incorporated through response prediction or brain-feature integration to improve model representations and downstream performance (Zhao et al., 2025; Moussa et al., 2025; Moussa & Toneva, 2025; Lu et al., 2026). These findings motivate using human brain activity to complement conventional visual supervision.

Human representations can guide visual learning through behavioral judgments as well as neural measurements. Muttenthaler et al. (2025) show that alignment with human similarity judgments across abstraction levels improves semantic organization and generalization. Extending this direction from overt perceptual judgments to neural responses, MRS (Chen et al., 2026) aligns brain and

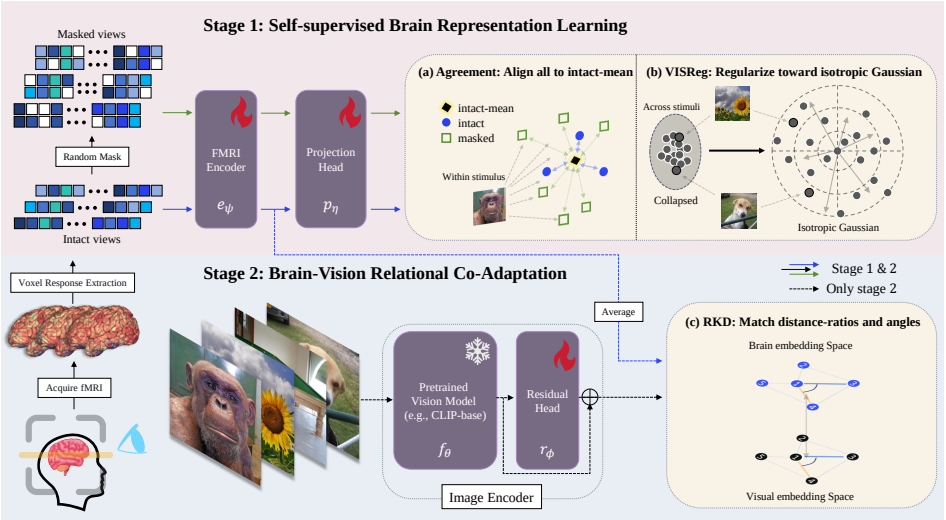

Figure 2: **Two-stage framework of Brain-VisA.** Stage 1 learns brain representations through agreement among views of each stimulus (a) and distributional regularization across stimuli (b); $T = 3$ is shown for illustration. (c) Stage 2 aligns brain and visual relations while retaining brain SSL.

visual representational structures through optimal transport (Peyré & Cuturi, 2019; Cuturi, 2013) and Gromov–Wasserstein matching (Peyré et al., 2016). Its use of shared category labels enables neural supervision to extend to a much larger image collection (Ridnik et al., 2021), but also introduces externally specified semantic grouping information. We instead study adaptation using only paired image–fMRI data, without category labels or additional training images.

3 Method

3.1 Problem Setup and Overview

Measurements and views. We model each participant separately. An *fMRI measurement* is a preprocessed voxel-response vector corresponding to one presentation of an image. For image x_i, let $v_{it} \in \mathbb{R}^D$ denote the t-th of T repeated measurements. The paired dataset is

$$\mathcal{D} = \left\{ \left(x_i, \{v_{it}\}_{t=1}^{T} \right) \right\}_{i=1}^{N}. \tag{1}$$

We call $\{v_{it}\}_{t=1}^{T}$ the *same-image fMRI group*. A *view* is an input constructed from a measurement: an intact view retains the original vector, whereas a masked view sets a randomly selected subset of voxel values to zero. Our goal is to adapt a pretrained visual model using these image–fMRI correspondences without semantic labels or category-conditioned image expansion.

Overview. Figure 2 summarizes Brain-VisA. Stage 1 learns stimulus-consistent brain representations through within-group agreement and across-image distributional regularization. Stage 2 co-adapts brain and visual representations through relational alignment while retaining brain SSL. A trainable Residual Head adapts a frozen Pretrained Vision Model, forming the *Image Encoder*.

3.2 Stage 1: Self-Supervised Brain Representation Learning

Cross-measurement agreement. The fixed stimulus in a same-image fMRI group anchors the learning objective to image-related information. For a minibatch of B images, each measurement v_{it} provides one intact view $v_{it}^{(0)} = v_{it}$ and two independently masked views $v_{it}^{(1)}$ and $v_{it}^{(2)}$. Thus, t indexes repeated measurements and $a \in \{0, 1, 2\}$ indexes the view constructed from each measurement. All $3T$ views of an image form its *same-image positive group*. A shared fMRI Encoder e_ψ and Projection Head p_η map them to

$$z_{it}^{(a)} = e_\psi(v_{it}^{(a)}) \in \mathbb{R}^d, \qquad q_{it}^{(a)} = p_\eta(z_{it}^{(a)}) \in \mathbb{R}^k. \tag{2}$$

Both modules are MLP-based, with the Projection Head following LeJEPA (Balestriero & LeCun, 2025); d matches the visual feature dimension. Every view projection is encouraged to agree with the mean intact-view projection of the same image (Figure 2(a)):

$$\mathcal{L}_{\text{pred}} = \frac{1}{3BTk} \sum_{i=1}^{B} \sum_{t=1}^{T} \sum_{a=0}^{2} \left\| q_{it}^{(a)} - \frac{1}{T} \sum_{t'=1}^{T} q_{it'}^{(0)} \right\|_2^2. \tag{3}$$

The mean target remains differentiable. This objective penalizes representational variation within a same-image positive group, encouraging consistency across repeated measurements and masking without reconstructing voxel values or separating negative pairs.

Across-image distributional regularization. Agreement alone allows collapsed representations. VISReg (Wu et al., 2026) instead encourages an isotropic Gaussian distribution of projected features across images (Figure 2(b)). For each (t, a), stack $Q_t^{(a)} = [q_{1t}^{(a)}, \ldots, q_{Bt}^{(a)}]^\top \in \mathbb{R}^{B \times k}$, with one image per row. The regularizer $\mathcal{R}_{\text{VIS}}$ controls the center, scale, and projected distribution of these features. The Stage 1 objective is

$$\mathcal{L}_{\text{SSL}} = \mathcal{L}_{\text{pred}} + \frac{\gamma}{3T} \sum_{t=1}^{T} \sum_{a=0}^{2} \mathcal{R}_{\text{VIS}}(Q_t^{(a)}). \tag{4}$$

Here γ weights regularization. $\mathcal{L}_{\text{SSL}}$ acts on projected features q while Stage 2 aligns encoder features z. Appendix A provides its full expression and the architecture and alignment details.

3.3 Stage 2: Brain–Vision Relational Co-Adaptation

Brain and visual representations. We initialize the fMRI Encoder and Projection Head from Stage 1. For image x_i, we average the encoder outputs of its intact measurements and adapt the frozen visual features with a Residual Head:

$$b_i = \frac{1}{T} \sum_{t=1}^{T} e_\psi(v_{it}), \qquad h_i = f_\theta(x_i) + r_\phi(f_\theta(x_i)). \tag{5}$$

Motivated by the frozen-backbone adaptation principle used in parameter-efficient methods such as LoRA (Hu et al., 2022), we learn an output-space residual correction (He et al., 2016) while keeping the Pretrained Vision Model f_θ fixed. This preserves the pretrained feature extractor while allowing brain supervision to reshape its output representations. The Residual Head r_ϕ is a lightweight MLP with a zero-initialized final layer, so the Image Encoder initially reproduces the pretrained visual features and subsequently learns a brain-guided correction.

Relational alignment. RKD (Park et al., 2019) aligns the relations among images in the two spaces (Figure 2(c)). Its loss $\mathcal{L}_{\text{align}}$ matches pairwise distances normalized by the mean distance in each space, together with angles formed by corresponding triplets. This transfers relative geometry rather than requiring individual brain and visual features to coincide.

Joint optimization and inference. Stage 2 retains brain SSL:

$$\mathcal{L}_{\text{stage 2}} = \mathcal{L}_{\text{SSL}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}}. \tag{6}$$

The SSL loss updates e_ψ and p_η, while the alignment loss updates e_ψ and r_ϕ. Continued SSL maintains the within-image agreement and across-image regularization as the two modalities co-adapt. At inference, only the Image Encoder is used; no fMRI is required.

4 Experiments

We first examine the semantic structure learned from fMRI, then test whether brain-guided adaptation improves visual abstraction and generalization. Controlled SSL comparisons and design variants identify which learning choices support these benefits; the final comparison evaluates MRS under the same downstream tasks.

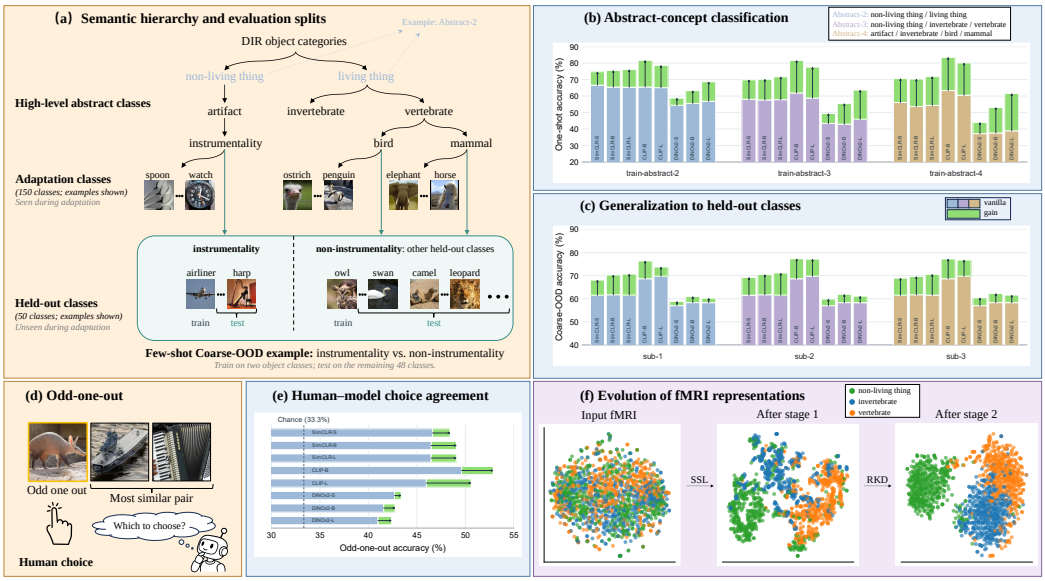

Figure 3: **Evaluation overview and gains over unadapted visual models.** (a) We uses WordNet-defined high-level classes for evaluation, while keeping adaptation and held-out object categories disjoint. (b,c,d,e) Brain-VisA improves high-level semantic abstraction and generalization, with gains across all 3 participants, and better matches human similarity judgments. (f) Illustrative brain-space t-SNE views at each stage and high-level semantic structure emerges through Brain-VisA.

4.1 Experimental Setup

Datasets. DIR provides paired images and fMRI from three participants (Shen et al., 2019). For each participant, adaptation uses 150 object categories $\times$ 8 images $\times$ 5 repeated measurements: $1,200$ distinct images paired with $6,000$ fMRI measurements, corresponding to $T = 5$ in equation 1. Another 50 object categories are excluded from adaptation and used to evaluate category generalization. We train participant-specific models and additionally examine the two-stage design on NSD with MindEye preprocessing (Allen et al., 2022; Scotti et al., 2023).

Training and comparisons. We evaluate eight pretrained vision models from three families: SimCLR, CLIP, and DINOv2. Comparisons include unadapted models (Vanilla), MRS (Chen et al., 2026), and substitutions of our brain SSL objective with autoencoding (AE), denoising autoencoding (DAE), or contrastive learning (CL) (Hinton & Salakhutdinov, 2006; Vincent et al., 2008; Chen et al., 2020). These SSL alternatives retain the paired data and downstream relational alignment.

Evaluation. Figure 3(a) summarizes the semantic hierarchy and evaluation splits. To assess semantic abstraction, train-abstract-2/3/4 uses WordNet groupings (Fellbaum, 1998) to evaluate one-shot classification into two, three, or four high-level classes on the adaptation images. The corresponding abstract labels are used only for evaluation and never supervise brain-guided adaptation. To assess generalization, test-abstract-2/3 evaluates these distinctions on images from concrete categories excluded from adaptation. Coarse-OOD tests whether a few-shot classifier fitted to support categories generalizes to other categories within the held-out split, averaging the five binary semantic tasks with the largest category coverage. THINGS provides a complementary behavioral evaluation, measuring agreement with human odd-one-out judgments across $1,854$ object concepts (Hebart et al., 2019). Appendix B provides training configurations, controlled comparisons, evaluation protocols, and reporting conventions.

4.2 Semantic Abstraction and Object-Category Recognition

Semantic structure in brain representations. Figure 4 suggests that positive-view agreement reveals clearer high-level semantic structure than voxel reconstruction. Reconstruction preserves measured responses containing both stimulus-related information and measurement-specific varia-

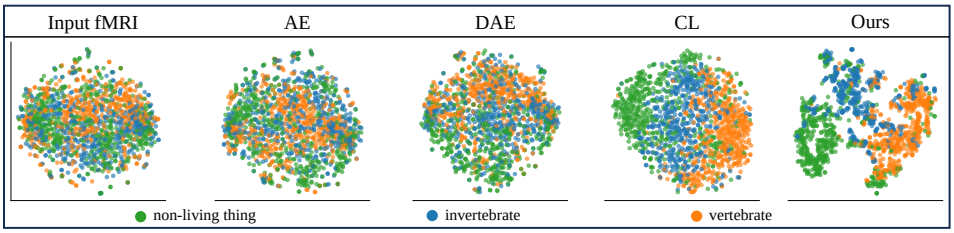

Figure 4: **Brain representation geometry under different self-supervised objectives.** t-SNE of input fMRI and Stage 1 encoder outputs before brain–vision alignment. AE and DAE prioritize reconstructing individual voxel responses and exhibit weaker semantic structure than CL and Brain-VisA, both of which encourage agreement among views of the same stimulus.

Table 1: Mean one-shot accuracy (%) over train-abstract-2/3/4. Results average 100 evaluation trials and, for adapted models, three DIR participants. The best result for each visual model is bold.

	SimCLR			CLIP		DINOv2		
Method	Small	Base	Large	Base	Large	Small	Base	Large
Vanilla	60.24	58.92	59.16	63.55	61.46	45.00	45.35	47.23
AE	66.49	66.33	65.18	67.05	58.91	48.17	51.42	56.48
DAE	64.66	64.40	63.22	62.36	56.96	48.61	52.53	57.68
CL	57.96	56.82	57.17	59.51	58.02	49.57	54.11	60.95
MRS	64.47	63.60	60.16	70.13	68.35	49.14	46.54	47.00
Brain-VisA	**71.73**	**72.02**	**73.18**	**82.24**	**78.74**	**50.73**	**57.27**	**64.61**

tion. Our brain SSL objective instead encourages agreement across repeated responses to the same image. Whereas masked views of one measurement share its trial-specific variability, separate measurements share the stimulus while varying across trials, encouraging agreement to capture stimulus-consistent information. The weaker grouping under Same-fMRI and the controlled transfer results in Section 4.5 further support the value of cross-measurement agreement.

High-level semantic abstraction in visual representations. Brain-guided adaptation improves abstract classification across all eight pretrained visual models. Figure 3(b) shows gains at each abstraction level, and Table 1 summarizes their mean. Averaged over tasks and models, Brain-VisA exceeds Vanilla and MRS by 13.70 and 10.14 percentage points, respectively. These gains are obtained using paired image–fMRI data alone, whereas MRS introduces category information to extend neural supervision beyond the measured pairs.

Preservation of concrete-category recognition. Table 2 examines whether improved semantic abstraction comes at the expense of distinguishing individual object categories. Alongside its abstraction gains, Brain-VisA largely preserves Vanilla accuracy on ImageNet-150 and ImageNet-1K under one-epoch linear evaluation, whereas MRS exhibits pronounced decreases. This pattern suggests that brain-guided adaptation can strengthen high-level semantic organization while retaining information useful for more specific category distinctions. Evaluation protocols, detailed classification results, and supplementary visualizations and training-cost comparisons are provided in Appendices B, C, and D, respectively.

4.3 Generalization Beyond the Adaptation Categories

The abstraction gains extend to object categories not used for brain-guided adaptation. Test-abstract-2/3 preserves the high-level classification question while replacing the concrete categories with the held-out DIR split. Brain-VisA exceeds Vanilla by 10.14 and 12.74 percentage points, respectively (Table 3). Thus, the improved semantic organization is not restricted to the object categories whose paired fMRI responses trained the model: it also supports the generalization to unseen categories.

The gains also extend to human similarity judgments and semantic OOD recognition. On THINGS, Brain-VisA increases mean agreement with human odd-one-out judgments from 45.05% to 47.32%

Table 2: Classification accuracy (%) averaged equally across eight visual models and, for adapted models, three DIR participants. Abstract-2/3/4 use one-shot classification on adaptation categories; ImageNet-150/1K use top-1 accuracy after one epoch of linear evaluation. The two column groups use different evaluation protocols. Best results per column are bold.

	Abstract classification			Concrete-category recognition	
Method	Abstract-2	Abstract-3	Abstract-4	ImageNet-150	ImageNet-1K
Vanilla	61.83	53.30	50.22	73.43	56.65
MRS	63.57	57.74	54.71	49.02	35.56
Brain-VisA	**72.16**	**67.42**	**66.86**	**73.69**	**56.68**

Table 3: Generalization accuracy (%), averaged across eight visual models and, for adapted models, three DIR participants. Test-abstract tasks evaluate one-shot abstract classification on held-out categories. Bold indicates the best mean for each metric.

Method	Test-abstract-2	Test-abstract-3	THINGS	Coarse-OOD
Vanilla	62.89	51.84	45.05	62.11
AE	66.37	56.56	45.94	64.45
DAE	65.43	55.33	45.79	63.87
CL	63.74	53.34	45.56	62.59
MRS	66.95	58.69	45.58	63.97
Brain-VisA	**73.03**	**64.58**	**47.32**	**68.12**

(Table 3), with improvements across all eight visual models (Figure 3(e)). On Coarse-OOD, mean accuracy increases from 62.11% to 68.12%, with gains observed for all three participants (Figure 3(c)). Together with the held-out abstract classification results, these findings support the benefits of brain-guided adaptation across category generalization and behavioral alignment. Appendices C.2 and D.3 report per-model results and the full semantic OOD evaluation.

4.4 Comparison of different SSL strategies

We compare our brain SSL objective with AE, DAE, and CL while keeping all other configurations fixed. Our method achieves the highest mean train-abstract-2/3/4 accuracy for every visual model (Table 1) and the highest cross-model mean on all four generalization metrics (Table 3). These results establish the advantage of our objective for both semantic abstraction and visual generalization.

The weaker transfer of AE and DAE is consistent with their less pronounced semantic organization after Stage 1 (Figure 4). CL introduces a different limitation, despite the emergence of high-level abstract semantic concepts during Stage 1. During Stage 2, the contrastive objective treats fMRI responses to different images as negatives and pushes their brain representations apart. Through relational alignment, this repulsion can indirectly push apart the visual representations of semantically similar images, disrupting semantic relationships in the visual model. Our objective combines agreement across repeated measurements with distributional regularization, avoiding explicit inter-image repulsion while providing more effective guidance for visual adaptation.

4.5 Analysis of different configurations

Cross-Measurement Positive Groups. This cross-measurement agreement yields clearer semantic organization in brain representations (Figure 7) and higher mean abstract classification accuracy across all three model families (Figure 5(a)). Views derived from a single measurement share its measurement-specific variation while grouping separate measurements instead anchors agreement to the fixed image and penalizes trial-specific representational differences, encouraging stimulus-consistent information to guide visual adaptation.

Benefits of Two-Stage Self-Supervision. On NSD, our method also improves THINGS performance across all three visual model families, supporting its applicability beyond DIR. We compare three variants: *Without* omits Stage 1 and uses only relational alignment in Stage 2; *Before* retains Stage 1 but removes SSL during Stage 2; and *With* skips Stage 1 while retaining both SSL and re-

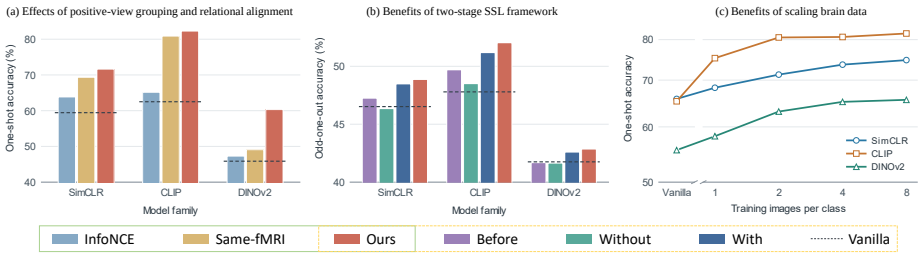

Figure 5: **Design choices and paired-data scale.** (a) DIR participant 3: mean train-abstract-2/3/4 accuracy. Within-stimulus grouping and relational alignment outperform single-measurement grouping and point-wise alignment, respectively. (b) NSD participant 2: THINGS accuracy. The full two-stage framework outperforms alternative training schedules. Both panels report averages within each model family. (c) DIR participant 3: train-abstract-2 accuracy versus images per category, with all five measurements retained. Increasing the training data size further improves performance.

lational alignment in Stage 2. Brain-VisA achieves the highest family-mean accuracy in all three families (Figure 5(b)). The weaker performance of *Without* supports the importance of brain SSL for extracting stable stimulus-related information from noisy fMRI measurements. The advantage over *Before* suggests that this constraint remains valuable during alignment: continued SSL regularizes the brain encoder, helping keep joint adaptation anchored to stimulus-consistent structure rather than allowing both modalities to evolve solely toward agreement with each other. The advantage over *With* supports the value of Stage 1 initialization, which provides a structured brain representation as the starting point for alignment. Together, these comparisons support complementary roles for the two stages: SSL pretraining establishes a useful brain representation, while continued SSL helps maintain its stimulus consistency during relational co-adaptation.

Relational versus Instance-Level Alignment. We replace RKD in Stage 2 with cross-modal InfoNCE, which attracts matched brain–image pairs and repels mismatched pairs, while retaining brain SSL. Figure 5(a) shows that relational alignment outperforms point-wise alignment with InfoNCE in high-level semantic abstraction across all three model families. InfoNCE emphasizes matching individual brain–image pairs, potentially losing relational information among stimuli. Preserving this representational structure is central to transferring the conceptual organization encoded in brain representations to vision models.

Effect of Paired-Data Scale. We vary the number of training images per category from one to eight while retaining all five fMRI measurements per image. Figure 5(c) shows an overall increase in abstract classification accuracy across the three model families, with gains already present at the smallest evaluated data scale. More distinct paired stimuli can therefore further benefit the learned representation beyond exploiting repeated measurements of a fixed stimulus.

This study characterizes the observed data-scale trend. Appendices B.3, C.3, and D.4 provide the comparison protocols, per-model results, and alignment-weight sensitivity analysis.

5 Conclusion

The proposed Brain-VisA identifies self-supervised brain representation learning as a key component of brain-guided visual alignment. Our two-stage framework combines agreement across repeated fMRI measurements with distributional regularization, followed by relational co-adaptation of brain and visual representations. Comparisons with alternative SSL objectives and controlled variants show that Brain-VisA learns more effective stimulus-consistent brain representations before and during visual adaptation. Across eight pretrained vision models, Brain-VisA improves high-level semantic abstraction and generalization, while largely preserving their original performance on fine-grained category recognition. The resulting image encoder requires only images at inference, allowing brain-derived supervision to benefit downstream visual tasks without further fMRI acquisition. Our current evaluation relies on participant-specific models; learning shared brain representations across participants remains an important direction for future work.

AI USE STATEMENT

We used generative AI tools to assist with language polishing and improve the clarity of the manuscript. The authors reviewed and revised the resulting text and take responsibility for the final content.

ETHICS STATEMENT

All data used in this study come from publicly available datasets, and no new human participant data were collected. We are not aware of additional ethical concerns arising from the analyses presented.

REPRODUCIBILITY STATEMENT

The main text and appendices document the datasets, model architectures, training configurations, and evaluation protocols, including key hyperparameters and reporting conventions. These details are provided to facilitate reproduction of the experiments and reuse of the proposed framework.

REFERENCES

Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, jan 2022. ISSN 1546-1726. doi: 10.1038/s41593-021-00962-x. URL https://doi.org/10.1038/s41593-021-00962-x.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Human-like cognitive generalization for large models via mental representation-guided supervision. *Nature Communications*, 17(1):4709, apr 2026. ISSN 2041-1723. doi: 10.1038/s41467-026-71267-5. URL https://doi.org/10.1038/s41467-026-71267-5.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15745–15753, 2021. doi: 10.1109/CVPR46437.2021.01549.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

Joel Dapello, Kohitij Kar, Martin Schrimpf, Robert Baldwin Geary, Michael Ferguson, David Daniel Cox, and James J. DiCarlo. Aligning model and macaque inferior temporal cortex representations improves model-to-human behavioral alignment and adversarial robustness. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SMYdcXjJh1q.

Callie Federer, Haoyan Xu, Alona Fyshe, and Joel Zylberberg. Improved object recognition using neural networks trained to mimic the brain's statistical properties. *Neural Networks*, 131:103–114, 2020. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2020.07.013. URL https://www.sciencedirect.com/science/article/pii/S0893608020302549.

Christiane Fellbaum (ed.). *WordNet: An Electronic Lexical Database*. The MIT Press, 05 1998. ISBN 9780262272551. doi: 10.7551/mitpress/7287.001.0001. URL https://doi.org/10.7551/mitpress/7287.001.0001.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, 2022. ISSN 1546-1726. doi: 10.1038/s41593-022-01026-4. URL https://doi.org/10.1038/s41593-022-01026-4.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent a new approach to self-supervised learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10):1–24, 10 2019. doi: 10.1371/journal.pone.0223792. URL https://doi.org/10.1371/journal.pone.0223792.

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. doi: 10.1126/science.1127647. URL https://www.science.org/doi/abs/10.1126/science.1127647.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Zhe Li, Wieland Brendel, Edgar Walker, Erick Cobos, Taliah Muhammad, Jacob Reimer, Matthias Bethge, Fabian Sinz, Zachary Pitkow, and Andreas Tolias. Learning from brains how to regularize machines. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/70117ee3c0b15a2950f1e82a215e812b-Paper.pdf.

Zitong Lu, Yile Wang, and Julie D. Golomb. Achieving more human brain-like vision via human eeg representational alignment. *Communications Biology*, 9(1):463, feb 2026. ISSN 2399-3642. doi: 10.1038/s42003-026-09685-w. URL https://doi.org/10.1038/s42003-026-09685-w.

Omer Moussa and Mariya Toneva. Brain-tuning improves generalizability and efficiency of brain alignment in speech models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=4jgsUhWWaF.

Omer Moussa, Dietrich Klakow, and Mariya Toneva. Improving semantic understanding in speech language models via brain-tuning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KL8Sm4xRn7.

Lukas Muttenthaler, Klaus Greff, Frieda Born, Bernhard Spitzer, Simon Kornblith, Michael C. Mozer, Klaus-Robert Müller, Thomas Unterthiner, and Andrew K. Lampinen. Aligning machine and human visual representations across abstraction levels. *Nature*, 647(8089):349–355, nov 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09631-6. URL https://doi.org/10.1038/s41586-025-09631-6.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 3967–3976. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00409. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Park_Relational_Knowledge_Distillation_CVPR_2019_paper.html.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Found. Trends Mach. Learn.*, 11(5–6):355–607, February 2019. ISSN 1935-8237. doi: 10.1561/2200000073. URL https://doi.org/10.1561/2200000073.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2664–2672, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/peyre16.html.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik. Imagenet-21k pretraining for the masses. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/98f13708210194c475687be6106a3b84-Paper-round1.pdf.

Shahd Safarani, Arne Nix, Konstantin Willeke, Santiago A. Cadena, Kelli Restivo, George Denfield, Andreas S. Tolias, and Fabian H. Sinz. Towards robust vision by multi-task learning on monkey visual cortex. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Martin Schrimpf, Jonas Kubilius, Michael J. Lee, N. Apurva Ratan Murty, Robert Ajemian, and James J. DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423, 2020. ISSN 0896-6273. doi: 10.1016/j.neuron.2020.07.040. URL https://doi.org/10.1016/j.neuron.2020.07.040.

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. doi: 10.1073/pnas.2105646118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2105646118.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, ethan cohen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, and Tanishq Abraham. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 24705–24728. Curran Associates, Inc., 2023. doi: 10.52202/075280-1073. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4ddab70bf41ffe5d423840644d3357f4-Paper-Conference.pdf.

Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLOS Computational Biology*, 15(1):1–23, 01 2019. doi: 10.1371/journal.pcbi.1006633. URL https://doi.org/10.1371/journal.pcbi.1006633.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 1096–1103, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390294. URL https://doi.org/10.1145/1390156.1390294.

Haiyu Wu, Randall Balestriero, and Morgan Levine. Visreg: Variance-invariance-sketching regularization for jepa training. *arXiv preprint arXiv:2606.02572*, 2026.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html.

Shuangchen Zhao, Changde Du, Jingze Li, Hui Li, and Huiguang He. Neuralood: Improving out-of-distribution generalization performance with brain-machine fusion learning framework. *Information Fusion*, 119:103021, 2025. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2025.103021. URL https://www.sciencedirect.com/science/article/pii/S1566253525000946.

A METHOD DETAILS

A.1 NETWORK ARCHITECTURES

The fMRI Encoder maps participant-specific voxel inputs to the visual feature dimension d. Its input layer is a linear map to $1,024$ units followed by GELU and LayerNorm. Four residual MLP blocks follow, each containing two $1,024$-unit linear layers, GELU, dropout 0.1, a skip connection, and LayerNorm. A final linear layer and LayerNorm produce the d-dimensional encoder output.

The Projection Head follows LeJEPA (Balestriero & LeCun, 2025): $d \to 2,048 \to 2,048 \to \lfloor d/3 \rfloor$, with BatchNorm and GELU after each of the first two layers. The visual Residual Head is $d \to 256 \to d$, with GELU, no dropout, and a zero-initialized final layer. Its output is added to the frozen pretrained visual feature. AE and DAE use an additional decoder, $d \to 1,024 \to D$, where D is the participant's voxel dimension; its hidden layer uses GELU and LayerNorm, without dropout.

A.2 BRAIN SELF-SUPERVISION AND VISREG

For a generic Q, let μ and σ be its coordinate-wise batch mean and population standard deviation, and let $\widetilde{Q}$ contain its centered features divided column-wise by $\mathrm{sg}(\sigma)$. The stop-gradient operator sg applies only to this standardization. Then

$$\mathcal{R}_{\mathrm{VIS}}(Q) = \underbrace{\frac{1}{k}\|\mu\|_2^2}_{\text{center}} + \underbrace{\frac{1}{k}\|\mathbf{1}_k - \sigma\|_2^2}_{\text{scale}} + \underbrace{\frac{1}{BM}\sum_{\ell=1}^{M} \|\operatorname{sort}(\widetilde{Q}w_\ell) - g_{\mathcal{N}}\|_2^2}_{\text{shape}}. \tag{7}$$

The three terms encourage zero mean, unit coordinate-wise standard deviation, and Gaussian-shaped projections along M random unit directions w_ℓ. For each measurement index t and view index a, the rows of $Q_t^{(a)} \in \mathbb{R}^{B\times k}$ correspond to B different images. For a generic Q, define

$$\mu_j = \frac{1}{B}\sum_{i=1}^{B} Q_{ij}, \qquad \sigma_j = \left[\frac{1}{B}\sum_{i=1}^{B}(Q_{ij}-\mu_j)^2\right]^{1/2}, \qquad \widetilde{Q} = (Q - \mathbf{1}_B\mu^\top)/\,\mathrm{sg}(\sigma), \tag{8}$$

where division is column-wise. Gradients are stopped through σ in this standardization, but not in the scale penalty in equation 7. For each of M random unit directions $w_\ell \in \mathbb{R}^k$, the shape term sorts $\widetilde{Q}w_\ell$ and matches it to

$$(g_{\mathcal{N}})_j = \Phi^{-1}\left(\frac{j}{B+1}\right), \qquad j = 1,\dots,B, \tag{9}$$

where Φ^{-1} is the standard Gaussian quantile function. The center and scale terms control coordinate-wise statistics; the shape term constrains projected distributions. The regularizer is averaged over all $3T$ measurement–view combinations as in equation 4.

We use $M = 256$ random projection directions and $\gamma = 1.5$. The center, scale, and shape terms each have unit internal weight. The implementation adds 10^{-6} to the standard deviation for numerical stability. Each intact measurement has two independently masked views, with each voxel zeroed with probability 0.3. The mean intact-view target remains differentiable.

A.3 RELATIONAL ALIGNMENT AND OPTIMIZATION

The features $\{b_i\}_{i=1}^B$ and $\{h_i\}_{i=1}^B$ represent the same images. For $u \in \{b, h\}$, let $\bar{d}_u$ denote the mean nonzero pairwise distance, and let $\widehat{\Delta}_{ij}^u = (u_j - u_i)/\|u_j - u_i\|_2$ for nonzero displacements.

Zero displacements have zero direction by convention. The RKD objective is

$$\mathcal{L}_{\text{align}} = \underbrace{\frac{1}{B^2}\sum_{i,j=1}^{B} \ell\Big(\frac{\|b_i - b_j\|_2}{\bar{d}_b}, \frac{\|h_i - h_j\|_2}{\bar{d}_h}\Big)}_{\text{distance ratios}} + \underbrace{\frac{2}{B^3}\sum_{i,j,l=1}^{B} \ell\Big(\langle \widehat{\Delta}^b_{ij}, \widehat{\Delta}^b_{il}\rangle, \langle \widehat{\Delta}^h_{ij}, \widehat{\Delta}^h_{il}\rangle\Big)}_{\text{angles}}, \tag{10}$$

where ℓ is the Smooth L1 penalty. The distance term compares each pair relative to the mean distance of its own space; the angle term compares corresponding triplet-angle cosines. The relative weight of 2 follows the RKD implementation. Both feature spaces receive gradients from $\mathcal{L}_{\text{align}}$.

The distance and angle terms have weights 1 and 2. Stage 2 jointly updates the fMRI Encoder and visual Residual Head through relational alignment; brain SSL also updates the fMRI Encoder and Projection Head. The visual backbone remains frozen, and only the adapted visual representation is needed at inference.

B Experimental Settings and Evaluation Protocols

B.1 Datasets, Preprocessing, and Visual Models

DIR uses participant-specific voxel spaces from the union of LOC, FFA, and PPA (HVC). Each participant contributes 150 adaptation categories with eight images and five measurements per image: $1,200$ images and $6,000$ measurements. Measurements are kept separately in positive groups. Another 50 categories are held out from adaptation. We use the stored preprocessed voxel responses and do not introduce additional normalization in the training data loader. The NSD experiment uses participant 2 with MindEye-preprocessed nsdgeneral voxels ($14,278$ dimensions), three measurements per image, and $8,700$ image presentations per training epoch.

Both datasets use eight visual models: SimCLR-small/base/large, CLIP-base/large, and DINOv2-small/base/large. Their output dimensions are $384/768/1,024$, $512/768$, and $384/768/1,024$, respectively. CLIP uses ViT-B/16 and ViT-L/14; SimCLR uses ViT-S/B/L with 16-pixel patches, and DINOv2 uses its small/base/large models. Images are center-cropped to 224×224 after resizing; the resize is 224 for SimCLR/CLIP and 256 for DINOv2. The corresponding backbone normalization is applied when extracting visual features. Held-out categories refer to the brain-guided adaptation split, not to the original pretraining corpora.

B.2 Training Configurations and Hyperparameters

Table 4 gives the common settings for the full two-stage runs. DIR uses all available adaptation images without an adaptation validation split. Main Brain-VisA results use the Stage 2 epoch-100 checkpoint; MRS uses epoch 20. A single training seed is used, and participants are trained independently. Evaluation trials are resampled support sets, not independent training runs.

Table 4: Training settings for Brain-VisA. Unless explicitly varied in a controlled comparison, the same settings apply to DIR and NSD.

Setting	Value
Optimizer	AdamW
Learning rate / weight decay	10^{-5} / 10^{-5}
Image batch size / random seed	100 / 42
Stage 1 / Stage 2 epochs	100 / 100
Masked views per measurement / zero-mask probability	2 / 0.3
Brain SSL weight in Stage 2 / VISReg weight γ	1 / 1.5
Alignment weight λ_{align}	10
RKD distance / angle weights	1 / 2
VISReg random projections	256
Encoder hidden width / residual blocks / dropout	$1,024$ / 4 / 0.1
Visual head hidden width / dropout	256 / 0

B.3 Baselines and Controlled Comparisons

Brain SSL alternatives. AE, DAE, CL, and Brain-VisA use the same paired images, participants, frozen backbone, fMRI encoder architecture, visual head, two-stage epoch schedule, optimizer, and RKD alignment settings. The brain objective and its required auxiliary module change. AE reconstructs intact voxels with mean squared error; DAE reconstructs intact voxels from independently zero-masked inputs with mask probability 0.3. These reconstruction baselines use the decoder in Appendix A.1 and do not add VISReg. CL uses multi-view contrastive brain SSL with temperature 0.1, treating same-image views as positives and different images as negatives. CL denotes brain SSL and is distinct from the cross-modal InfoNCE comparison. Brain-VisA uses agreement plus VISReg as in equation 4.

Positive groups and cross-modal alignment. DIR participant 3 is used for both comparisons in Figure 5(a). Same-fMRI keeps all repeated measurements and their masked views, but computes agreement separately within each measurement's three-view group. Brain-VisA groups all $3T$ same-image views. Data, architecture, masking, optimizer, and alignment remain fixed. The InfoNCE

variant keeps brain SSL and replaces RKD with symmetric cross-modal contrastive alignment (temperature 0.07), with matched image–brain pairs as positives. The alignment weight remains 10; this holds the coefficient fixed, not the numerical scale of the different losses.

SSL schedules on NSD. Participant 2 and all eight visual models are used. Without omits Stage 1 and uses RKD alone in Stage 2; Before uses Stage 1 SSL and RKD alone in Stage 2; With omits Stage 1 and uses both SSL and RKD in Stage 2; Brain-VisA uses both stages. Stage 2 lasts 100 epochs in every condition, and Stage 1 lasts 100 epochs when present. The dataset, architecture, optimizer, and alignment weight are fixed. Total training exposure differs when Stage 1 is omitted, so this is a schedule comparison rather than a total-compute-matched comparison.

Paired-data scale. Figure 5(c) uses DIR participant 3 with 150 categories. Each data subset selects $K \in \{1, 2, 4, 8\}$ images per category and retains all five repeated measurements, giving $150K$ images and $750K$ measurements. Stage 1 uses 100 epochs for each subset. Stage 2 is evaluated at epochs 800, 400, 200, and 100, respectively, so each configuration processes $120,000$ image presentations during Stage 2. With a batch size of 100 and retained partial batches, $K = 1$ requires $1,600$ Stage 2 updates, while the other settings require $1,200$. Stage 1 updates also vary with subset size. The experiment therefore matches Stage 2 image exposure, not total computation or total optimization steps.

The plot reports train-abstract-2 accuracy averaged over models within each family: three SimCLR sizes, two CLIP sizes, and three DINOv2 sizes. Evaluation uses the full adaptation-category image pool rather than shrinking with the training subset. Vanilla denotes the unadapted representation and is a separate reference, not an additional paired-data scale. The curves describe the measured data-scale trend and are not a fitted scaling law.

Alignment-weight sensitivity. The DIR participant-2 sweep changes only $\lambda_{\text{align}} \in \{1, 5, 10, 20, 30\}$, retaining RKD, both SSL stages, VISReg, the full adaptation data, and the other settings in Table 4. All values are evaluated at Stage 2 epoch 100. The $\lambda = 10$ column is the standard full-data run; it is not an additional independently trained replicate.

MRS and Vanilla. Vanilla uses the unadapted backbone representation. MRS uses the same visual backbones, DIR participants, and downstream evaluation tasks, with its own training procedure and epoch-20 checkpoint. MRS extends supervision to auxiliary images using category information, whereas Brain-VisA uses the $1,200$ measured DIR images. This is a comparison of complete methods, not a controlled substitution of the brain SSL objective. Parameter and timing accounting is described in Appendix D.5.

B.4 Downstream Evaluation Protocols

Abstract classification. Train-abstract-2/3/4 uses the adaptation-category image pool; test-abstract-2/3 uses the held-out category pool. The WordNet groupings are living/non-living for two classes, invertebrate/vertebrate/non-living for three, and bird/invertebrate/mammal/artifact for four, according to the task-specific category mappings. Only categories covered by a grouping are evaluated: train-abstract-2/3/4 covers $150/146/137$ concrete categories; test-abstract-2/3 covers all 50 held-out categories. Each trial samples one support image per abstract class and evaluates the remaining covered images, excluding the sampled supports. We average 100 trials. Features are L2-normalized, and logistic regression uses L2 regularization with $C = 1$, the L-BFGS solver, and up to $5,000$ iterations. The stored train-abstract results use the float64 GPU backend and seed 42. Held-out abstract and OOD evaluations use the scikit-learn implementation; logged Brain-VisA evaluations use seed $42 + e$ at epoch e. These evaluation labels never supervise the brain-guided adaptation model.

Semantic OOD. Each of 31 binary tasks divides the 50 held-out concrete categories into a target semantic group and its complement. A trial samples one concrete category from each side and five support images from each selected category. The classifier is tested on images from all other concrete categories, excluding both support categories. It uses the same feature normalization and logistic-regression settings as above, with 100 trials per task. We report ordinary accuracy. Coarse-OOD averages the five tasks with the largest category coverage: device, vertebrate, living thing, instrumentality, and non-living thing, covering 8, 11, 15, 23, and 35 held-out categories. Table 5 lists all task coverages. Groupings overlap across tasks, so their coverage counts are not additive.

THINGS. We use one designated image per concept for all $1,854$ concepts and evaluate human odd-one-out triplets with cosine similarity between visual representations. The predicted odd image is the image outside the most similar pair. Agreement is pooled over the supplied training and validation triplet files; these names denote the behavioral data partitions, and no classifier is fitted to the judgments. Ties are counted as correct when the human-selected pair attains the highest similarity.

ImageNet retention. We use concrete-category recognition as a supplementary retention test. A separate linear classifier is trained for one epoch on each frozen representation, with matched classifier initialization across Vanilla, MRS, and Brain-VisA. ImageNet-1K uses the official validation set. ImageNet-150 uses the DIR adaptation categories and a custom split: the first 50 images per category in filename order form a $7,500$-image validation set, and the remaining images train the classifier. Evaluation labels train only the evaluation classifier, not the brain-guided adaptation model. The scores characterize this one-epoch protocol rather than fully converged linear evaluation. The linear classifiers use SGD with learning rate 0.01 and momentum 0.9; the evaluation scripts specify training/evaluation batch sizes of $1,280/2,560$ and seed 42.

Table 5: Concrete-class coverage of the 31 semantic OOD tasks in the DIR test and adaptation splits. Counts indicate classes, not images. Bold rows marked with † are the five tasks used for Coarse-OOD: they have the largest test-class coverage and also the largest training-class coverage. Coarse-OOD averages their accuracies with equal weights. Targets can overlap or have ancestor–descendant relationships, so counts must not be summed across rows. The non-living category is the complement of the living category within each split.

OOD task	Test classes (50)	Train classes (150)
insect	2	3
percussion_instrument	2	1
protective_covering	2	0
home_appliance	2	3
box	2	1
hat	2	0
craft	2	7
self_propelled_vehicle	2	2
chordophone	2	0
aquatic_bird	2	4
bird	3	6
ungulate	3	4
headdress	3	1
equipment	3	16
arthropod	3	8
invertebrate	4	12
structure	4	5
wheeled_vehicle	4	6
stringed_instrument	4	1
clothing	4	3
musical_instrument	5	4
consumer_goods	6	6
mammal	6	16
covering	6	4
vehicle	6	13
container	7	17
device†	**8**	**45**
vertebrate†	**11**	**27**
living_thing†	**15**	**43**
instrumentality†	**23**	**91**
non_living_thing†	**35**	**107**

B.5 Reporting, Aggregation, and Visualization

Main adapted-model DIR results average participants 1–3; Vanilla features do not depend on a participant. Table 1 first averages train-abstract-2/3/4 for each model, and Table 3 averages the eight models separately for each metric. The main design figure averages small/base/large within SimCLR and DINOv2 and base/large within CLIP in both panels (a) and (b). Controlled studies and sensitivity analysis use the specified single participant. Metrics are not combined into an overall generalization score. Accuracies are percentages and gains are absolute percentage-point differences, computed from unrounded values before display to two decimal places.

The available DIR brain-embedding visualizations use L2-normalized representations, PCA to 50 dimensions, and t-SNE with perplexity 50, PCA initialization, automatic learning rate, $1,000$ iterations, and random seed 42. Balanced measurement subsets contain $3,440$, $1,440$, or 960 points for abstract-2/3/4. A point represents one measurement, not an average image response. Layouts are fitted separately, so absolute coordinates across panels are not directly comparable. Figure 4 compares Stage 1 brain encoder outputs before brain–vision alignment. All panels in Figure 7 use the same participant. Quantitative claims about the learning objectives and positive-group construction are supported by the controlled performance comparisons.

C Detailed Experimental Results

C.1 Abstract Semantic Classification

Tables 6–8 separate the three tasks summarized in Table 1, with results for each visual model. Table 9 additionally separates the MRS and Brain-VisA results by participant.

Table 6: One-shot accuracy (%) on train-abstract-2, averaged over 100 evaluation trials and, for adapted models, three DIR participants. CL denotes contrastive brain self-supervision; MRS denotes the mental representation-guided baseline. Best results per model are bold.

	SimCLR			CLIP		DINOv2		
Method	Small	Base	Large	Base	Large	Small	Base	Large
Vanilla	66.60	65.44	65.34	65.47	65.10	54.42	55.47	56.82
AE	71.20	71.12	69.67	69.13	63.95	56.84	59.61	63.17
DAE	69.92	69.61	68.25	65.76	62.82	57.13	60.21	63.76
CL	65.06	63.86	63.82	63.48	63.45	57.82	61.09	66.05
MRS	69.20	67.78	64.64	69.07	67.13	**59.02**	56.04	55.70
Brain-VisA	**74.75**	**75.52**	**76.06**	**81.65**	**78.60**	58.76	**63.29**	**68.66**

Table 7: One-shot accuracy (%) on train-abstract-3. Evaluation settings and abbreviations follow Table 6. Best results per model are bold.

	SimCLR			CLIP		DINOv2		
Method	Small	Base	Large	Base	Large	Small	Base	Large
Vanilla	58.04	57.62	57.81	61.84	58.76	43.36	42.99	45.96
AE	63.86	64.30	63.41	64.86	55.77	46.66	49.28	55.24
DAE	61.85	62.17	60.99	59.54	53.60	47.13	50.48	56.56
CL	54.85	54.32	54.74	56.88	55.16	48.16	52.09	59.86
MRS	62.19	62.49	59.95	71.28	68.27	47.75	44.96	45.04
Brain-VisA	**69.83**	**70.20**	**71.67**	**81.61**	**77.55**	**49.43**	**55.45**	**63.62**

Table 8: One-shot accuracy (%) on train-abstract-4. Evaluation settings and abbreviations follow Table 6. Best results per model are bold.

	SimCLR			CLIP		DINOv2		
Method	Small	Base	Large	Base	Large	Small	Base	Large
Vanilla	56.08	53.71	54.33	63.34	60.51	37.23	37.60	38.92
AE	64.41	63.57	62.47	67.17	57.02	41.02	45.38	51.04
DAE	62.23	61.42	60.44	61.78	54.47	41.57	46.91	52.70
CL	53.97	52.28	52.94	58.18	55.47	42.73	49.14	56.94
MRS	62.02	60.53	55.89	70.04	69.65	40.66	38.60	40.27
Brain-VisA	**70.60**	**70.33**	**71.80**	**83.46**	**80.09**	**44.01**	**53.08**	**61.55**

Table 9: Participant-specific mean one-shot accuracy (%) over train-abstract-2/3/4. Columns 1–3 denote DIR participants. Values are averaged from the unrounded task scores; Vanilla is shared across participants. These results underlie the MRS and Brain-VisA rows of Table 1.

Model	Vanilla	MRS 1	MRS 2	MRS 3	Brain-VisA 1	Brain-VisA 2	Brain-VisA 3
SimCLR-small	60.24	63.89	64.35	65.18	70.95	72.59	71.64
SimCLR-base	58.92	63.91	62.77	64.13	72.71	72.63	70.71
SimCLR-large	59.16	60.65	60.33	59.50	73.71	73.29	72.54
CLIP-base	63.55	69.27	71.29	69.83	80.62	83.22	82.87
CLIP-large	61.46	68.86	67.70	68.49	71.88	82.67	81.68
DINOv2-small	45.00	49.49	49.15	48.79	48.18	51.56	52.45
DINOv2-base	45.35	46.58	47.02	46.01	52.59	59.05	60.17
DINOv2-large	47.23	46.74	47.76	46.51	59.22	66.30	68.30

C.2 GENERALIZATION

Tables 10–13 give the per-model results underlying Table 3. Adapted results average three DIR participants.

Table 10: One-shot accuracy (%) on test-abstract-2. Evaluation settings and abbreviations follow Table 6. Best results per model are bold.

	SimCLR			CLIP		DINOv2		
Method	Small	Base	Large	Base	Large	Small	Base	Large
Vanilla	67.97	68.43	67.07	64.93	61.29	55.20	59.15	59.08
AE	70.98	72.82	71.45	69.97	62.93	58.39	62.38	62.05
DAE	69.88	71.60	70.16	66.03	61.71	58.69	62.87	62.49
CL	65.25	65.79	65.86	64.00	62.58	59.36	63.67	63.43
MRS	69.84	71.65	68.72	75.87	71.17	58.26	59.09	61.01
Brain-VisA	**75.03**	**77.81**	**78.18**	**85.21**	**78.87**	**59.79**	**64.71**	**64.62**

Table 11: One-shot accuracy (%) on test-abstract-3. Evaluation settings and abbreviations follow Table 6. Best results per model are bold.

	SimCLR			CLIP		DINOv2		
Method	Small	Base	Large	Base	Large	Small	Base	Large
Vanilla	55.23	54.12	53.79	60.19	56.02	43.29	45.84	46.26
AE	64.88	63.94	62.82	62.15	53.70	46.14	50.38	48.44
DAE	63.81	62.69	61.68	56.52	51.46	46.43	50.96	49.09
CL	58.30	55.95	56.83	53.80	52.58	47.09	51.87	50.30
MRS	64.25	66.99	63.80	69.11	66.11	46.81	45.71	46.71
Brain-VisA	**69.50**	**69.29**	**70.25**	**79.94**	**75.21**	**47.51**	**53.08**	**51.89**

Table 12: Accuracy (%) on the THINGS odd-one-out evaluation. Best results per model are bold.

	SimCLR			CLIP		DINOv2		
Method	Small	Base	Large	Base	Large	Small	Base	Large
Vanilla	46.61	46.48	46.47	49.61	45.98	42.70	41.58	40.99
AE	47.55	47.92	47.57	50.38	47.02	43.13	42.22	41.71
DAE	47.34	47.71	47.35	49.70	46.86	43.19	42.32	41.82
CL	46.53	46.66	46.66	49.40	47.16	43.37	42.58	42.11
MRS	**48.54**	48.64	48.01	51.39	48.76	41.17	39.23	38.91
Brain-VisA	48.43	**49.08**	**49.06**	**52.88**	**50.59**	**43.38**	**42.75**	**42.37**

Table 13: Accuracy (%) on the Coarse-OOD evaluation, averaged over the five selected semantic OOD tasks. Best results per model are bold.

	SimCLR			CLIP		DINOv2		
Method	Small	Base	Large	Base	Large	Small	Base	Large
Vanilla	61.54	61.73	61.59	68.67	69.76	57.01	58.32	58.29
AE	65.62	66.57	65.74	70.23	67.96	58.92	60.70	59.87
DAE	64.83	65.80	64.98	68.26	66.95	59.09	60.95	60.09
CL	61.41	62.06	62.23	66.72	67.13	59.42	61.29	60.48
MRS	64.82	65.25	63.33	71.69	70.69	58.34	58.43	59.19
Brain-VisA	**68.75**	**70.14**	**70.80**	**76.97**	**75.90**	**59.71**	**61.72**	**60.98**

C.3 Design Comparisons

Figure 6 separates the eight visual models underlying the family means in Figure 5. Panel (a) uses DIR participant 3, and panel (b) uses NSD participant 2. Both include SimCLR-small/base/large, CLIP-base/large, and DINOv2-small/base/large; no six-model subset is used. The controls and schedule definitions are given in Appendix B.3.

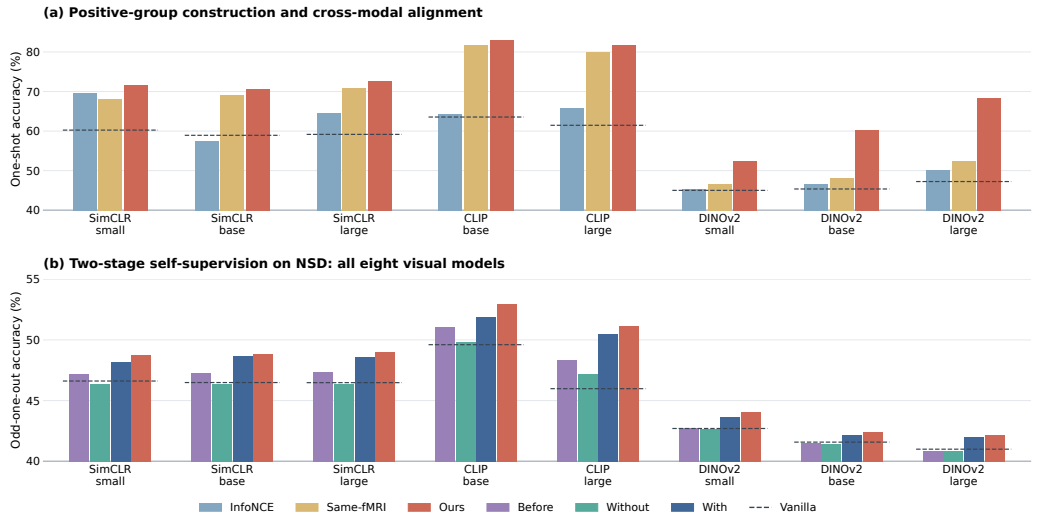

Figure 6: **Per-model design comparisons.** (a) DIR participant 3: mean train-abstract-2/3/4 accuracy. (b) NSD participant 2: THINGS accuracy. Both panels contain all eight visual models used in the main family averages.

C.4 IMAGENET EVALUATION

Tables 14 and 15 report per-model results, averaged over three DIR participants, for the one-epoch retention evaluation summarized in Table 2. The classifier protocol is described in Appendix B.4.

Table 14: ImageNet-1K validation top-1 accuracy (%) under one-epoch linear probing. A separate linear classifier is trained on each frozen representation using the same classifier initialization. MRS and Brain-VisA use the epoch-20 and Stage-2 epoch-100 checkpoints, respectively, averaged over three DIR participants. Best results per model are bold, determined before rounding.

	SimCLR			CLIP		DINOv2		
Method	Small	Base	Large	Base	Large	Small	Base	Large
Vanilla	19.54	36.25	35.99	**50.63**	**72.04**	**75.73**	80.43	82.60
MRS	7.50	17.37	19.69	11.22	35.41	55.68	67.06	70.55
Brain-VisA	**19.77**	**36.25**	**36.07**	50.60	72.00	75.71	**80.43**	**82.60**

Table 15: ImageNet-150 top-1 accuracy (%) under one-epoch linear probing. The 150 concrete categories correspond to the DIR adaptation classes. For each class, the first 50 images in filename order are held out for validation (7,500 images in total); the remaining images train the linear classifier. This is a custom split, not a subset of the official ImageNet-1K validation split. Representations remain frozen, and the three classifiers share the same initialization. Adapted-model results average three DIR participants. Best results per model are bold, determined before rounding.

	SimCLR			CLIP		DINOv2		
Method	Small	Base	Large	Base	Large	Small	Base	Large
Vanilla	34.59	59.12	57.23	65.79	**90.93**	91.12	93.76	**94.89**
MRS	11.73	28.88	31.93	13.09	40.32	83.21	90.64	92.36
Brain-VisA	**35.24**	**59.34**	**57.34**	**67.08**	90.68	**91.16**	**93.80**	**94.89**

D SUPPLEMENTARY ANALYSES

D.1 POSITIVE-GROUP VISUALIZATION

Figure 7 complements the main-text discussion of cross-measurement agreement. Same-fMRI retains all measurements but forms separate three-view groups; Brain-VisA forms one $3T$-view group per image. On DIR participant 3, Brain-VisA exceeds Same-fMRI by 3.84, 5.70, and 6.72 percentage points on train-abstract-2/3/4, respectively, averaged across the eight visual models. The visualization is qualitative; the controlled performance comparison supplies the quantitative evidence.

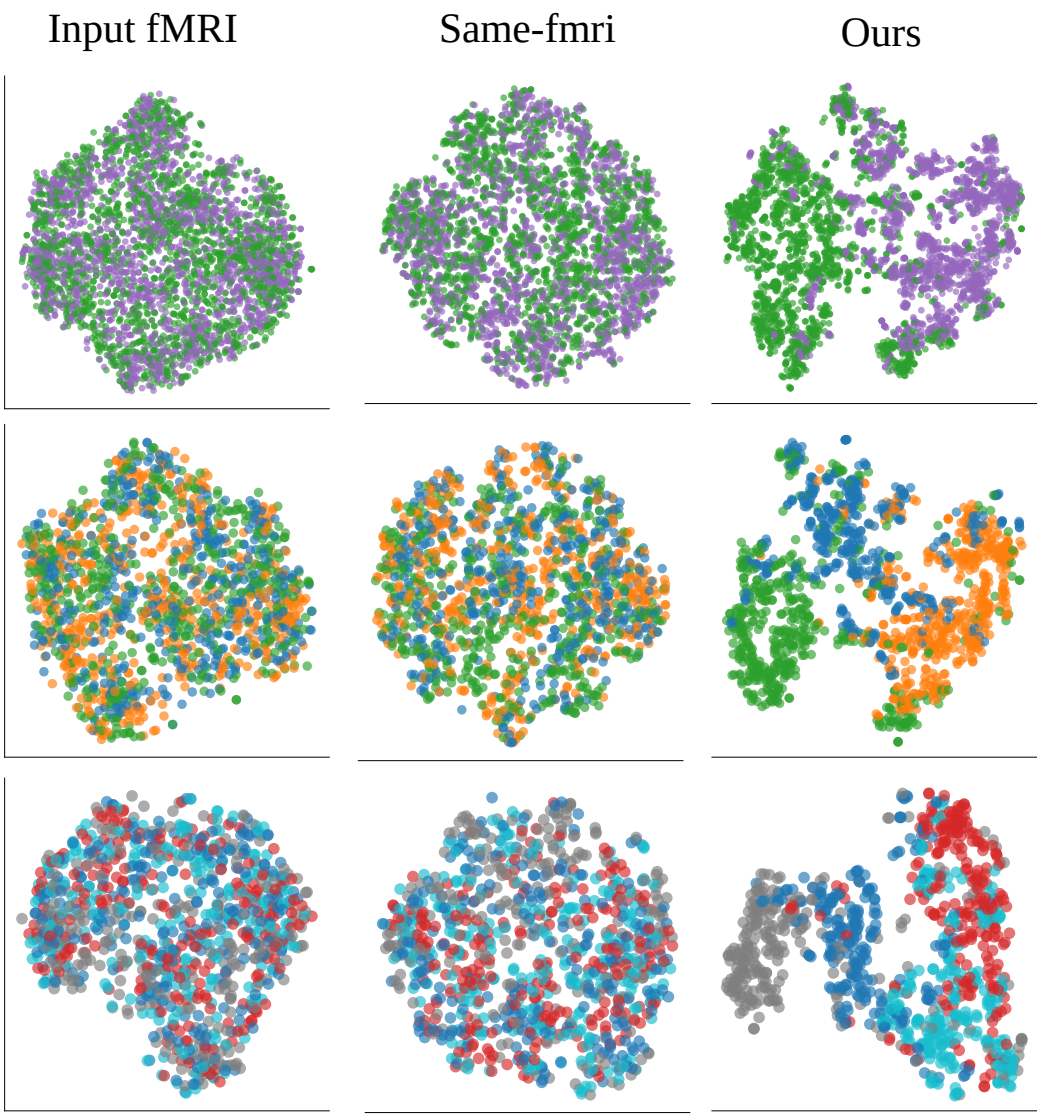

Figure 7: **Brain representations under different positive groups.** t-SNE of input fMRI and learned brain representations from the same participant. Each point is one measurement; layouts are fitted separately. Quantitative results are given in Figure 5(a).

D.2 SSL VISUALIZATION

Figure 8 shows that across the three abstraction settings, abstract-2/3/4, input fMRI and reconstruction-based representations show substantial overlap between abstract categories, while contrastive learning reveals partial separation. Brain-VisA exhibits more pronounced category-related structure, consistent with the semantic abstraction gains reported in the main text. These visualizations provide a qualitative view of how different self-supervised objectives shape brain representations before brain–vision alignment.

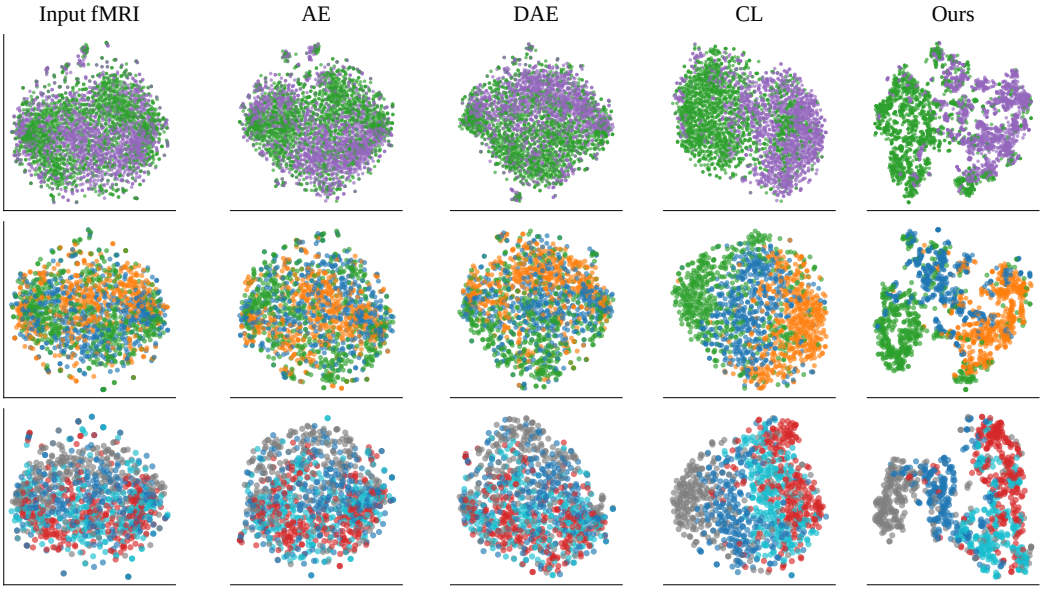

Figure 8: **Semantic structure in brain representations under different self-supervised objectives.** t-SNE visualizations of input fMRI and Stage 1 encoder outputs from the same participant. Rows correspond to Abstract-2, Abstract-3, and Abstract-4, from top to bottom. Each point represents an fMRI measurement, colored by its abstract category. The t-SNE layout is fitted separately for each representation.

D.3 EVALUATION ACROSS ALL SEMANTIC OOD TASKS

Table 16 reports the equal-weight average over all 31 binary semantic OOD tasks. The five-task Coarse-OOD summary emphasizes broadly populated semantic groups; the full-task average also includes narrower groups. Task definitions and coverage are given in Appendix B.4 and Table 5.

Table 16: Mean accuracy (%) across all 31 binary semantic OOD tasks. Results average 100 evaluation trials per task and, for adapted models, three DIR participants. Each task receives equal weight. CL denotes contrastive brain self-supervision; MRS denotes the mental representation-guided baseline. Best results per model are bold.

	SimCLR			CLIP		DINOv2		
Method	Small	Base	Large	Base	Large	Small	Base	Large
Vanilla	52.71	52.33	52.29	54.63	54.46	50.06	49.21	48.69
AE	53.68	53.40	53.35	55.40	54.19	52.28	52.06	50.26
DAE	53.65	53.35	53.31	55.16	53.98	52.31	52.09	50.30
CL	53.33	52.88	52.91	55.09	54.16	52.38	52.16	50.40
MRS	53.58	53.20	53.57	55.24	**55.60**	51.52	50.55	50.42
Brain-VisA	**53.89**	**53.86**	**53.84**	**56.03**	55.35	**52.42**	**52.23**	**50.52**

D.4 Alignment-Weight Sensitivity

Table 17 reports the mean of train-abstract-2/3/4 for each model on DIR participant 2. These means are calculated from the unrounded task-level outputs. All tested weights improve on Vanilla, but the preferred range differs across visual families: SimCLR performs best around 5–10, CLIP around 10–20, and DINOv2 at 1 among the evaluated weights. Thus, the common setting $\lambda_{\text{align}} = 10$ is not optimal for every backbone. The sweep reports sensitivity of the fixed main configuration; it does not replace the main results with model-specific tuned settings.

Table 17: Mean one-shot accuracy (%) over train-abstract-2/3/4 on DIR participant 2 as the alignment weight varies. All other training settings are fixed. Bold marks the highest unrounded adapted score within each model; the final row averages the eight models.

Model	Vanilla	$\lambda = 1$	5	10	20	30
SimCLR-small	60.24	67.00	**73.12**	72.59	71.36	70.54
SimCLR-base	58.92	65.99	**72.99**	72.63	71.88	71.32
SimCLR-large	59.16	66.36	**73.29**	73.29	72.76	72.38
CLIP-base	63.55	75.57	82.70	83.22	**83.32**	83.23
CLIP-large	61.46	74.09	81.75	82.67	**82.74**	82.51
DINOv2-small	45.00	**57.06**	52.76	51.56	49.94	48.87
DINOv2-base	45.35	**64.90**	60.55	59.05	56.80	54.80
DINOv2-large	47.23	**68.86**	66.83	66.30	65.22	63.65
Mean	55.12	67.48	70.50	70.17	69.25	68.41

D.5 Data Usage, Model Size, and Training Cost

Although Brain-VisA uses 100 epochs in each of its two stages, compared with 20 epochs for MRS, it achieves stronger semantic abstraction and generalization in our main evaluations and has a lower recorded training cost. Epoch counts alone do not capture this difference: MRS uses category labels to expand the adaptation image pool from 1,200 to over 200,000 images, substantially increasing the amount of data processed per epoch, whereas Brain-VisA uses only the 1,200 measured images per participant. MRS also uses Wasserstein and Gromov–Wasserstein alignment with iterative numerical optimization, including Sinkhorn updates and nested transport optimization. These operations introduce computational overhead that the direct distance- and angle-based relational losses in Brain-VisA avoid. The parameter difference additionally reflects the architectures: MRS uses a transformer-based brain encoder and a GNN, while Brain-VisA uses an MLP-based brain encoder without a GNN. Together, these differences help explain the lower training cost and parameter count in the evaluated configurations.

Table 18 counts active trainable modules, excluding frozen visual backbones. The timing estimates in Table 19 describe the recorded configurations rather than a matched-hardware, matched-budget speed benchmark. In particular, Brain-VisA excludes evaluation and checkpointing intervals, and the alignment weights used for timing differ from the main DIR setting. These estimates support a configuration-specific training-cost comparison, not an intrinsic algorithmic speedup factor.

Table 18: Total trainable parameters for DIR participant 1 using HVC (LOC, FFA, and PPA), in millions. MRS includes the fMRI encoder, GNN, and image encoder; Brain-VisA includes the fMRI encoder, Projection Head, and Image Encoder. Frozen visual backbones and inactive MRS retrieval branches are excluded. Lower parameter counts are bold.

Model	Feature dim.	MRS (M)	Brain-VisA (M)	MRS/Brain-VisA
SimCLR-small	384	27.699	**19.658**	1.41×
SimCLR-base	768	105.138	**21.298**	4.94×
SimCLR-large	1024	185.273	**22.390**	8.27×
CLIP-base	512	47.810	**20.203**	2.37×
CLIP-large	768	105.138	**21.298**	4.94×
DINOv2-small	384	27.699	**19.658**	1.41×
DINOv2-base	768	105.138	**21.298**	4.94×
DINOv2-large	1024	185.273	**22.390**	8.27×

Table 19: Recorded training-time estimates in hours for DIR participant 2. MRS uses 20 epochs; Brain-VisA uses 100 epochs in each stage. Brain-VisA excludes evaluation and checkpointing intervals. MRS timings are measured for four representative feature dimensions and reused for models with the same dimension. Brain-VisA timings use alignment weight 1 for SimCLR models and 50 for CLIP and DINOv2 models. These configuration-specific measurements are not a matched-setting speed benchmark. Lower total times are bold; ratios use unrounded values.

	MRS	Brain-VisA			
Model	Total 20 epochs	Stage 1 100 epochs	Stage 2 100 epochs	Total 200 epochs	MRS/Brain-VisA
SimCLR-small	28.10	0.571	0.720	**1.290**	21.78×
SimCLR-base	28.51	0.803	0.530	**1.334**	21.38×
SimCLR-large	28.23	0.840	0.530	**1.370**	20.60×
CLIP-base	27.78	0.917	0.673	**1.590**	17.47×
CLIP-large	28.51	0.898	0.522	**1.420**	20.07×
DINOv2-small	28.10	0.900	0.793	**1.693**	16.60×
DINOv2-base	28.51	0.972	0.614	**1.586**	17.98×
DINOv2-large	28.23	1.002	0.556	**1.559**	18.11×