TRAJECTORY-INDUCED MODES: DECOMPOSING LAYERWISE REPRESENTATION FLOW IN DEEP NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks progressively transform internal representations across depth, yet how representational structure persists through this process remains poorly understood. Existing analyses characterize layerwise content, cross-layer similarity, or output decodability, but do not directly characterize persistence as a property of representation flow. We introduce Trajectory-Induced Modes (TIMs), which models stimulus-specific hidden-state trajectories through a reduced spectral operator and uses eigenvalue magnitude as an operational measure of cross-layer persistence, distinguishing attenuated, preserved, and amplified components. Across 108 frozen pretrained models spanning vision, audio, and language, TIMs reveals prevalent persistent modes with systematic modality-dependent profiles. Language models are dominated by highly persistent components, vision models exhibit broader mixtures of attenuation and amplification, and audio models show an intermediate pattern. Crucially, TIM persistence is not merely a spectral property. Higher-persistence components retain substantially more information aligned with terminal representations, and removing them from intermediate activations produces larger task-loss increases than removing transient or random TIM-subspace components. Higher persistence is also associated with stronger cortical alignment in audio and language, with more limited gains in vision. These findings remain consistent across perturbation, layer subsampling, rank variation, and residual-structure controls. Together, these results establish cross-layer persistence as a measurable and functionally meaningful property of deep representations, offering a new perspective on how information is transformed across network depth.

1 INTRODUCTION

Deep neural networks progressively transform representations across layers, preserving some components while attenuating or amplifying others. Yet, whether these transformations exhibit a systematic organization of cross-layer persistence remains unclear. Existing methods have examined what individual layers encode (Alain & Bengio, 2016; Bricken et al., 2023; Cunningham et al., 2024), how representations differ between layers (Raghu et al., 2017; Kornblith et al., 2019; Kriegeskorte et al., 2008), how information changes with depth (Shwartz-Ziv & Tishby, 2017), and when intermediate states become predictive of network outputs (nostalgebraist, 2020; Belrose et al., 2023). These approaches provide complementary views of layerwise computation, but do not directly quantify how individual representational components persist, attenuate, or amplify across depth. Consequently, whether deep networks exhibit a systematic organization of cross-layer persistence, and whether such organization carries functional significance, remains insufficiently understood.

Biological sensory systems provide a useful motivation for studying persistence. Along the visual hierarchy, low-level features are progressively transformed into object-level representations, while auditory processing moves from acoustic detail toward phonological and semantic information (DiCarlo & Cox, 2007; DiCarlo et al., 2012; Hickok & Poeppel, 2007). These transformations motivate two questions for deep neural networks: can we systematically quantify the persistence of individual representational components across depth, and does this persistence relate to subsequent computation?

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.

We introduce Trajectory-Induced Modes (TIMs), a spectral framework that operationalizes cross-layer persistence as a measurable property of stimulus-specific representation trajectories. Treating network depth as an ordered progression of hidden states, TIMs fits a trajectory-restricted affine surrogate and decomposes its reduced operator on the observed trajectory subspace. The resulting eigenvalue magnitudes provide operational persistence scores that distinguish attenuated, persistent, and amplified spectral modes. These scores characterize the fitted transition rather than directly measuring semantic invariance or assuming globally linear network dynamics. TIMs further constructs persistence-conditioned representatives, enabling systematic evaluation of the relationship between spectral persistence, terminal-output retention, functional intervention, and cortical alignment.

We evaluate TIMs across 108 frozen pretrained models spanning vision, audio, and language, including CNN-based and transformer-based architectures. The resulting spectra reveal systematic, modality-dependent persistence profiles, with prevalent near-unit modes and distinct patterns of attenuation and amplification. These profiles remain robust to perturbation, layer subsampling, and rank variation, while held-out transition prediction confirms that the fitted surrogate captures ordered cross-layer structure. Beyond establishing the existence and robustness of this organization, we investigate its functional significance. Terminal-held-out analyses show that persistent representatives retain substantially greater alignment with unseen terminal representations, while projection ablations across nine models demonstrate larger task-loss increases when persistent rather than transient components are removed. These effects also exceed random TIM-subspace controls in eight of the nine models. Finally, fMRI encoding analyses reveal persistence-dependent cortical alignment across vision, audio, and language, providing an external characterization of the relationship between spectral persistence and biological responses.

Together, these findings establish cross-layer persistence as a measurable and systematically organized property of deep network representations, linked to subsequent representational retention, functional sensitivity, and modality-dependent cortical correspondence.

Our main contributions are:

- We formulate cross-layer persistence as a stimulus-specific representation-flow problem, establishing a component-level perspective for characterizing how representations persist, attenuate, or amplify across network depth.
- We develop Trajectory-Induced Modes (TIMs), a trajectory-restricted spectral framework that operationalizes cross-layer persistence through reduced-operator eigenvalues and constructs persistence-conditioned representatives for downstream evaluation.
- We reveal systematic and robust cross-layer persistence organization across 108 models spanning vision, audio, and language, and demonstrate its functional significance through stronger terminal-output retention, greater task-loss impact under intervention, and enhanced cortical alignment.

2 Related Work

2.1 Representation Analysis Across Layers

Existing representation analyses largely focus on layer-level content, cross-layer geometry, information-theoretic summaries, or output decodability. Probing classifiers (Alain & Bengio, 2016), attribution methods (Sundararajan et al., 2017), circuit analyses (Olah et al., 2020; Elhage et al., 2021), and sparse autoencoders (Bricken et al., 2023; Cunningham et al., 2024) identify semantic or functional features, neurons, and circuits. Cross-layer methods such as SVCCA (Raghu et al., 2017), CKA (Kornblith et al., 2019), representational similarity analysis (Kriegeskorte et al., 2008), intrinsic dimensionality, and low-rank analyses (Ansuini et al., 2019) quantify changes in layer geometry. Information-theoretic analyses such as the Information Bottleneck (Shwartz-Ziv & Tishby, 2017) track mutual information with inputs and labels across depth, but summarize layers with aggregate quantities rather than identifying which components are preserved or lost. Output-oriented methods such as the logit lens (nostalgebraist, 2020) and tuned lens (Belrose et al., 2023) examine when intermediate states become predictive of final outputs. These approaches reveal content, geometry, information summaries, or decodability, but do not directly characterize individual components by their persistence across successive layers.

2.2 Dynamical and Operator-Theoretic Views

The dynamical-systems perspective on deep networks (Chen et al., 2018; Haber & Ruthotto, 2018) views depth as continuous time or discretized state evolution, mainly for architecture design, stability analysis, and continuous-depth modeling. Related operator-theoretic tools, including Dynamic Mode Decomposition (Schmid, 2010; Brunton et al., 2016) and Koopman analysis (Mezić, 2005; Dogra & Redman, 2020), decompose evolving signals into modes associated with persistence, decay, or growth under an estimated transition. TIMs extends this mode-based perspective to feedforward inference, where the ordered sequence is the stimulus-specific chain of layerwise activations rather than a physical time series. Its reduced operator characterizes only the observed trajectory subspace, while retention and intervention experiments assess whether the extracted persistence structure is relevant to subsequent computation.

2.3 Information Processing as State Evolution

A representation-flow view is also connected to broader perspectives that treat information processing as an evolving state-dependent process. In neuroscience, population activity is commonly studied as trajectories through state space (Vyas et al., 2020), where neural dynamics support perception, memory, and action. Cortical responses can also maintain stimulus-related information beyond the immediate input (King et al., 2016), suggesting that neural processing is not only a pointwise response to momentary stimulation. This state-based view also appears in JEPA-based world models (LeCun, 2022; Assran et al., 2023), where compact latent representations are preserved or predicted to support reasoning over latent dynamics rather than raw observations alone (Ha & Schmidhuber, 2018). These perspectives motivate examining how representational components propagate through the finite sequence of layers during a single feedforward pass.

3 Method

TIMs quantifies cross-layer persistence by treating the representations induced by each input across network depth as a finite trajectory. It fits a trajectory-restricted affine surrogate to summarize the observed layerwise transitions and decomposes the resulting reduced spectral operator to identify attenuated, persistent, and amplified modes. These modes are then used to construct persistence-conditioned representatives for evaluating representational retention and cortical alignment.

3.1 Trajectory-Restricted Affine Surrogate

We fit a shared affine surrogate to each stimulus-specific trajectory to summarize its layerwise transitions. The surrogate provides a common operator for spectral analysis without assuming that individual layers share the same transformation or that network dynamics are globally linear.

Trajectory-level local affine structure. Let $x_1, \dots, x_L \in \mathbb{R}^d$ denote the representations of a fixed input across L sampled layers, forming a finite trajectory $\mathcal{T} = \{x_1, \dots, x_L\}$ with transitions $x_{\ell+1} = \mathcal{F}_\ell(x_\ell)$. A first-order expansion around each observed state gives

$$\mathcal{F}_\ell(x) \approx \mathcal{F}_\ell(x_\ell) + D\mathcal{F}_\ell(x_\ell)(x - x_\ell) = J_\ell x + b_\ell, \quad b_\ell := \mathcal{F}_\ell(x_\ell) - J_\ell x_\ell. \tag{1}$$

Here, $J_\ell = D\mathcal{F}_\ell(x_\ell)$ where differentiable, and b_ℓ is the local affine offset. These local operators may vary across layers.

Affine lifting and trajectory-level surrogate. To jointly estimate the linear component and affine offset, we augment each state as $\tilde{x}_\ell := [x_\ell^\top, 1]^\top \in \mathbb{R}^{d+1}$. The layer-specific affine map becomes $\tilde{x}_{\ell+1} \approx \tilde{J}_\ell \tilde{x}_\ell$, where $\tilde{J}_\ell = \left[\begin{smallmatrix} J_\ell & b_\ell \\ 0 & 1 \end{smallmatrix}\right]$. Defining $X_1 = [x_1, \dots, x_{L-1}]$, $X_2 = [x_2, \dots, x_L]$, and $\tilde{X}_1 = [\tilde{x}_1, \dots, \tilde{x}_{L-1}]$, we fit a shared affine surrogate to the observed transitions:

$$(A_*, c_*) = \arg\min_{A,c} \|X_2 - AX_1 - c\mathbf{1}^\top\|_F^2, \qquad [A_*\ c_*] = X_2 \tilde{X}_1^\dagger. \tag{2}$$

Here, $\dagger$ denotes the Moore–Penrose pseudoinverse. The lifted operator $\tilde{A}_* = \left[\begin{smallmatrix} A_* & c_* \\ 0 & 1 \end{smallmatrix}\right]$ gives the predictor $x_{\ell+1} \approx A_* x_\ell + c_*$.

Residual interpretation and scope. We assess the surrogate fit using $E_{\text{rel}} = \|X_2 - [A_* \; c_*]\tilde{X}_1\|_F / \|X_2\|_F$. This residual measures the error incurred by approximating layer-dependent transitions with a shared affine operator. Since $L \ll d$, we analyze the fitted operator on the observed trajectory subspace rather than interpreting it as a global model of network dynamics.

3.2 Reduced Spectral Operator and Trajectory-Induced Modes

We construct a reduced spectral operator to quantify component-level persistence and use its modes to form persistence-conditioned representatives.

Subspace projection and reduced operator. Since $L \ll d$, the fitted linear component A_* is identifiable only through its action on the trajectory subspace $\mathcal{S} := \mathrm{col}(X_1)$. We remove the fitted affine offset by defining $Y_2 := X_2 - c_*\mathbf{1}^\top$, such that $Y_2 \approx A_* X_1$. Using the compact SVD $X_1 = U_r \Sigma_r V_r^\top$, where r is the effective rank, we construct

$$\hat{A}_r := U_r^\top Y_2 V_r \Sigma_r^{-1} \in \mathbb{R}^{r \times r}. \tag{3}$$

This reduced operator represents the fitted linear transition on $\mathcal{S}$, excluding the affine offset from spectral decomposition.

Trajectory-Induced Modes. We eigendecompose $\hat{A}_r v_i = \lambda_i v_i$, obtaining eigenpairs $\{(\lambda_i, v_i)\}_{i=1}^r$, which define the Trajectory-Induced Modes (TIMs). Each eigenvalue magnitude $\rho_i := |\lambda_i|$ provides an operational persistence score under the fitted transition: values near one indicate approximate preservation, while values below or above one indicate attenuation or amplification, respectively. Persistence characterizes the spectral behavior of a trajectory component under the fitted operator, rather than the invariance of its semantic content or activation across individual layers. Whether persistent components retain information relevant to subsequent computation is examined empirically.

TIM-specific trajectory representatives. To evaluate persistence levels, we construct spectrally weighted representatives for each stimulus. For stimulus n, let $\{x_{\ell,n}\}_{\ell=1}^L$ denote its trajectory, with $X_{1,n} = [x_{1,n}, \ldots, x_{L-1,n}]$ and $X_{2,n} = [x_{2,n}, \ldots, x_{L,n}]$. After estimating $(A_{*,n}, c_{*,n})$, we define $Y_{2,n} := X_{2,n} - c_{*,n}\mathbf{1}^\top$. The reduced eigendecomposition yields eigenpairs $\{(\lambda_i, v_i)\}_{i=1}^r$, which are mapped to the original feature space as $\phi_i = Y_{2,n} V_r \Sigma_r^{-1} v_i \in \mathbb{C}^d$, forming the TIM basis $\Phi_n = [\phi_1, \ldots, \phi_r]$. For a target persistence γ, Gaussian weighting emphasizes modes with $|\lambda_i|$ near γ:

$$w_i^{(\gamma)} = \frac{\exp\big(-(|\lambda_i| - \gamma)^2/(2\sigma^2)\big)}{\sum_{j=1}^r \exp(-(|\lambda_j| - \gamma)^2/(2\sigma^2))}.$$

Here, σ controls the selection bandwidth. We compute the depth-averaged representation $\bar{x}_n = \frac{1}{L}\sum_{\ell=1}^L x_{\ell,n}$, estimate coefficients $\beta_n = \arg\min_\beta \|\bar{x}_n - \Phi_n \beta\|_2^2$, and reconstruct the representative as $z_n^{(\gamma)} = \mathrm{Re}\{\Phi_n(w^{(\gamma)} \odot \beta_n)\}$, where $\odot$ denotes element-wise multiplication and the real part accounts for complex eigenpairs. Across N stimuli, these representatives form $Z^{(\gamma)} = \{z_n^{(\gamma)}\}_{n=1}^N$.

3.3 Brain Alignment from TIM-Specific Representatives

We use persistence-conditioned representatives to assess whether cross-layer persistence corresponds to cortical responses. For each model m and persistence setting γ, fMRI stimuli indexed by trial or time segment yield $Z_m^{(\gamma)} = [z_m^{(\gamma)}(1), \ldots, z_m^{(\gamma)}(T)]$. Following each dataset's protocol, we align these representations to fMRI recordings through feature-wise HRF convolution (Friston et al., 1998) using stimulus timing and resample them at acquisition times, obtaining $\bar{z}_m^{(\gamma)}(t)$. Let $y_r(t)$ denote the measured response at ROI r. We fit a cross-validated ridge encoding model $f_{r,m}^{(\gamma)}$ from aligned TIM representations to neural responses, yielding $\hat{y}_{r,m}^{(\gamma)}(t) = f_{r,m}^{(\gamma)}(\bar{z}_m^{(\gamma)}(t))$. Brain alignment is evaluated on held-out data using squared Pearson correlation:

$$a_{r,m}^{(\gamma)} = \mathrm{corr}^2\left(\hat{y}_{r,m}^{(\gamma)}(t), y_r(t)\right), \qquad a_m^{(\gamma)} = (a_{1,m}^{(\gamma)}, \ldots, a_{R,m}^{(\gamma)}) \in \mathbb{R}^R.$$

ROI-level scores are aggregated across held-out runs and sessions to obtain group-level alignment profiles.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

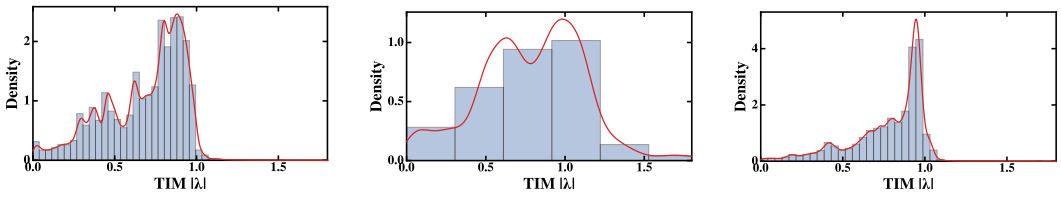
We evaluate TIMs across vision, audio, and language. Standalone spectral analyses use Caltech-256 (Griffin et al., 2007), ESC-50 (Piczak, 2015), and Penn Treebank (PTB) (Marcus et al., 1993). Brain-alignment analyses use THINGS-fMRI (Hebart et al., 2024), with images from THINGS (Hebart et al., 2023), the LeBel natural-language listening dataset (LeBel et al., 2023; 2025), and the Le Petit Prince multilingual corpus (Li et al., 2022; 2025). Following dataset-specific protocols, fMRI recordings are normalized to MNI152 space (Fonov et al., 2011), and voxel responses are averaged within Schaefer-200 cortical parcels (Schaefer et al., 2018). Causal interventions use WikiText-2 (Merity et al., 2017) for masked-language modeling, LibriSpeech validation-clean (Panayotov et al., 2015) for CTC speech recognition, and Imagenette (Howard, 2019) for image classification.

We analyze 108 frozen pretrained models (36 per modality), covering CNN-based and transformer-based vision models (ResNet (He et al., 2016), ViT (Dosovitskiy et al., 2021), DINOv2 (Oquab et al., 2024)), speech models (Wav2Vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), Whisper (Radford et al., 2023)), and language models (BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), T5 (Raffel et al., 2020)). Causal interventions use nine models: BERT-Base, RoBERTa-Base, and DistilBERT-Base; Wav2Vec2-Base, Data2Vec-Audio-Base, and UniSpeech-SAT-Base; and ViT-Base, DeiT-Base, and DINOv2-Base.

We apply consistent architecture-specific extraction rules across sampled layers: spatial pooling for CNNs, token pooling for ViTs and language models, and model-specific temporal pooling for audio. TIM statistics are computed independently within each model, without comparing raw feature coordinates across architectures.

4.2 DEEP NETWORKS EXHIBIT STRUCTURED PERSISTENCE ACROSS LAYERS

We first examine whether cross-layer persistence exhibits systematic organization across models and modalities. For each of the 108 models, we estimate the distribution of TIM eigenvalue magnitudes $|\lambda|$ using kernel density estimation (KDE). We classify modes as transient ($|\lambda| \in [0, 0.1)$), persistent ($|\lambda| \in [0.9, 1.1)$), or amplified ($|\lambda| \geq 1.2$).



(a) Audio (b) Vision (c) Language

Figure 1: TIM $|\lambda|$ distributions across modalities on the standalone stimulus datasets.

Figure 1 reveals persistence profiles across modalities. Persistent modes are more prevalent than transient modes in all three modalities ($p < 0.0001$ for each paired comparison), accounting for 42.5% of modes in language, 24.7% in vision, and 17.8% in audio (Table 1). Vision has the largest amplified fraction (13.2%), whereas amplification is rare in audio and language. Similar profiles on fMRI stimuli and Gaussian-random inputs suggest that these patterns are not specific to semantic stimulus content and may partly reflect architecture-dependent propagation (Appendices A and B).

4.3 TIM PERSISTENCE PREDICTS REPRESENTATIONAL RETENTION

Spectral persistence alone does not establish whether a component retains information relevant to subsequent computation. We therefore compare persistent (P, $\gamma = 1.0$), mid-persistence (M, $\gamma = 0.5$), and transient (T, $\gamma = 0.0$) representatives ($\sigma = 0.1$) using cosine similarity to the terminal representation. To prevent target leakage, TIM extraction, coefficient fitting, and representative

construction use only $x_1, \ldots, x_{L-1}$; x_L is reserved for evaluation. Additional in-trajectory analyses and per-model results appear in Appendices F and G.

Table 1: TIM mode prevalence and terminal-held-out output alignment. TIMs are constructed from $x_1, \ldots, x_{L-1}$ and evaluated against x_L; p-values use paired one-sided Wilcoxon tests.

Modality	KDE peak	Mode proportion			Terminal-held-out alignment				
		Transient	Persistent	Amplified	T	M	P	P−T	p-value
Audio	0.809	2.2%	17.8%	0.1%	-0.0427	-0.0707	0.2591	+0.3018	1.25×10^{-5}
Vision	0.957	3.3%	24.7%	13.2%	0.0668	0.1561	0.3517	+0.2849	2.86×10^{-4}
Language	0.906	1.0%	42.5%	0.0%	0.2407	0.2896	0.5146	+0.2739	1.47×10^{-6}

Table 1 shows significantly higher terminal-held-out alignment for persistent than transient representatives across all three modalities, with mean P–T differences ranging from $+0.2739$ to $+0.3018$. The persistent representatives constructed from preceding layers retain more structure aligned with the unseen terminal state, linking spectral persistence to subsequent representation without assuming that individual modes encode fixed semantic content. This is an aggregate tendency rather than a universal per-model ordering.

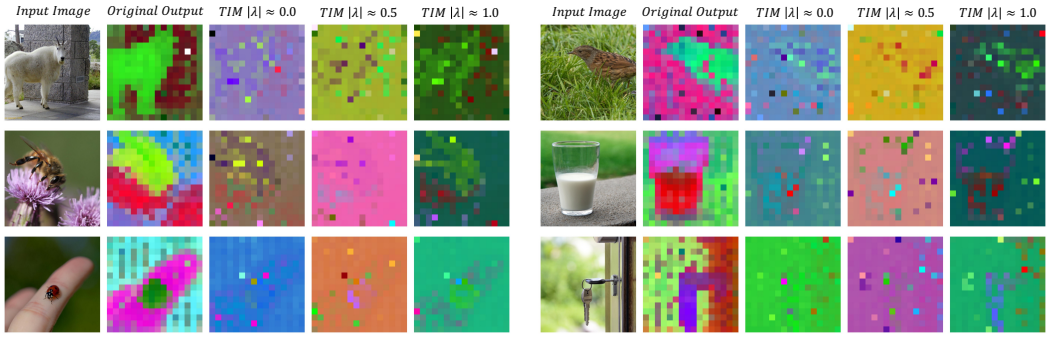


Figure 2: PCA visualization of TIM-reconstructed visual representations from DINOv2-Base (Oquab et al., 2024) under different $|\lambda|$-based persistence settings.

Figure 2 illustrates this pattern in DINOv2-Base: persistent reconstructions more closely resemble terminal-layer representation geometry, whereas lower-persistence reconstructions are more dispersed. Corresponding language and audio visualizations are provided in Appendix F.

4.4 Persistent Modes Are Functionally Consequential

Greater terminal-output alignment does not establish whether persistent components contribute to task performance. We therefore remove persistence-conditioned projections from intermediate activations and measure the resulting task-loss changes on disjoint evaluation examples. Gaussian weighting ($\sigma = 0.1$) yields unit-normalized persistent (d_P, $\gamma = 1$) and transient (d_T, $\gamma = 0$) directions from the calibrated TIM basis. At the middle transformer block, we intervene as

$$h'_{k,i} = h_{k,i} - \alpha \langle h_{k,i}, d_c \rangle d_c, \qquad c \in \{P, T, R\}, \tag{4}$$

where d_R is a random unit direction from the same TIM subspace. We intervene on the class token in vision and all token or time positions in language and audio, using $\alpha \in \{0.05, 0.10, 0.20\}$ and five random draws per control. Task-loss change is $\Delta\mathcal{L}_c = \mathcal{L}_c - \mathcal{L}_{\text{clean}}$. After averaging intervention strengths within each model, we apply paired one-sided Wilcoxon tests across nine models with Benjamini–Hochberg correction. We also normalize loss changes by the removed activation ratio.

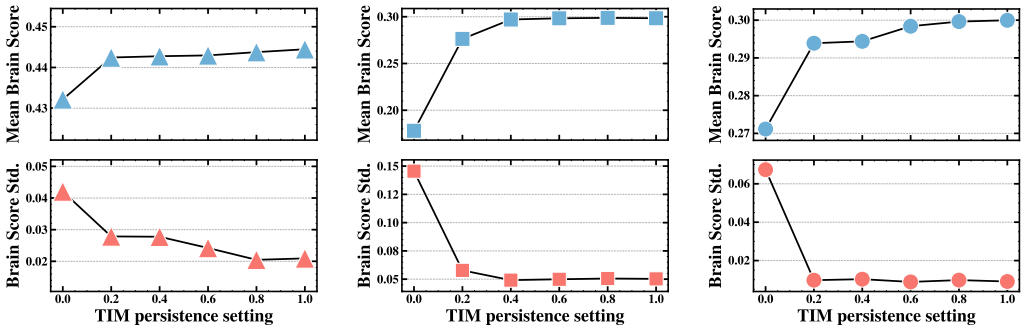
Persistent-component ablation produces larger task-loss increases than transient ablation in all nine models and exceeds the random TIM-subspace control in eight (Table 2). Both contrasts are significant (P–T: $q = 0.0039$; P–R: $q = 0.0098$), including after normalization by removed activation (P–T: $q = 0.0078$; P–R: $q = 0.0371$). Persistent-ablation effects increase with α in every model. These results show that the directions identified by high spectral persistence have greater functional impact under the tested interventions, extending the retention findings beyond representational similarity.

Table 2: Projection ablation of persistent (P), transient (T), and random TIM-subspace (R) components. Values are mean task-loss increases relative to clean inference.

Modality	$\alpha = 0.05$			$\alpha = 0.10$			$\alpha = 0.20$		
	$\Delta\mathcal{L}_P$	$\Delta\mathcal{L}_T$	$\Delta\mathcal{L}_R$	$\Delta\mathcal{L}_P$	$\Delta\mathcal{L}_T$	$\Delta\mathcal{L}_R$	$\Delta\mathcal{L}_P$	$\Delta\mathcal{L}_T$	$\Delta\mathcal{L}_R$
Language	0.0095	0.0028	0.0014	0.0220	0.0071	0.0042	0.0575	0.0205	0.0149
Audio	0.1368	-0.0026	0.0012	0.2806	-0.0051	0.0023	0.5037	-0.0103	0.0055
Vision	0.0010	-0.0002	0.0012	0.0024	-0.0004	0.0026	0.0063	-0.0010	0.0059
Overall significance: P>T, $q = 0.0039$; P>R, $q = 0.0098$.									

4.5 Cortical Alignment Depends on Persistence and Modality

We further examine whether cross-layer persistence corresponds to cortical responses. Using brain alignment as an external evaluation rather than an optimization target, we compare fMRI encoding performance across persistence settings γ and against alternative cross-layer aggregation methods.



(a) Audio (b) Vision (c) Language

Figure 3: Brain alignment as a function of the TIM persistence setting γ.

Figure 3 shows that mean cortical alignment increases with γ across all three modalities, while cross-run variability decreases at higher persistence settings.

Table 3: Brain alignment comparison between persistent-setting TIM representatives and alternative cross-layer aggregation methods. Scores are squared Pearson correlations; parentheses report paired one-sided Wilcoxon p-values against TIM.

Modality	**TIM**	**CKA-weighted**	**PCA aggregation**	**Uniform mean**
Audio	0.4445±0.0206	0.4355±0.0294 (< 0.0001)	0.3951±0.0229 (< 0.0001)	0.4354±0.0303 (< 0.0001)
Vision	0.2984±0.0496	0.2969±0.0447 (0.4598)	0.2929±0.0207 (0.0489)	0.2954±0.0465 (0.2253)
Language	0.2999±0.0102	0.2921±0.0090 (0.0062)	0.2298±0.0153 (< 0.0001)	0.2924±0.0086 (0.0239)

Persistent-setting TIM representatives show higher mean cortical alignment than alternative cross-layer aggregation methods across all three modalities (Table 3). The differences are statistically significant against CKA-weighted, PCA, and uniform-mean aggregation in audio and language. In vision, the difference is significant only against PCA aggregation. These comparisons provide complementary evidence for the relationship between spectral persistence and cortical correspondence, rather than establishing a universal performance advantage.

Figure 4 shows the anatomical distribution of cortical alignment, with ROI-level scores aggregated using a cross-model signal-to-noise ratio. Audio models exhibit alignment near temporal auditory regions, vision models over posterior and ventral visual areas, and language models across lateral temporal and frontal regions. These patterns suggest partial correspondence with modality-specific cortical processing without implying shared neural mechanisms.

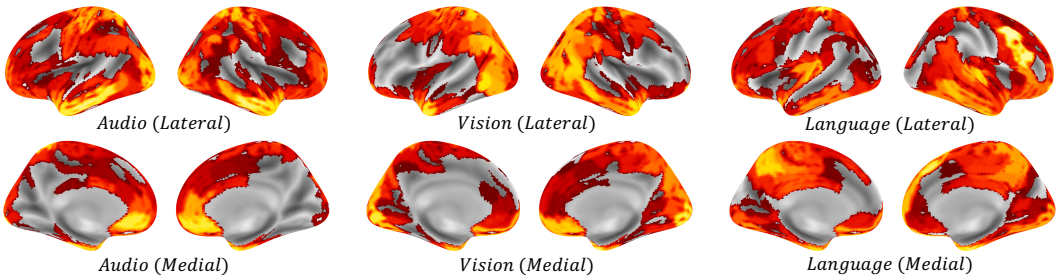


Figure 4: Cortical alignment maps for persistent-setting TIM representatives ($\gamma = 1.0$). Maps show ROI-level alignment scores for audio, vision, and language models.

4.6 Validation and Robustness of TIMs

We assess whether TIMs captures ordered cross-layer structure and whether the resulting persistence profiles depend on fitting or sampling choices. For the affine surrogate in Section 3.1, every third transition pair $(x_\ell, x_{\ell+1})$ is held out during fitting. We evaluate R^2 and relative error E_{rel} against a layer-shuffled baseline that preserves activations but disrupts their order. Mean held-out R^2 ranges from 0.893 to 0.931, with relative errors of 0.128–0.147 (Appendix C.1).

Table 4: Validation of the trajectory-restricted surrogate and stability of TIM extraction. Every third transition is excluded from fitting. Perturbation shift is measured at $\epsilon = 0.05$; layer-sampling statistics compare 75% sampling with the full-layer results.

Modality	Transition prediction		Perturbation	75% layer sampling		
	Held-out R^2	ΔR^2 vs. shuffle	Relative shift	Profile cosine	JS divergence	Mean-$\lvert\lambda\rvert$ ρ
Language	0.931	0.780	0.0131	0.913	0.053	0.701
Audio	0.918	0.910	0.0115	0.943	0.029	0.933
Vision	0.893	0.869	0.0202	0.949	0.022	0.972

The surrogate consistently outperforms the layer-shuffled baseline, indicating that it captures ordered cross-layer structure (Table 4). TIM extraction is also stable under moderate perturbation and layer subsampling. At $\epsilon = 0.05$, relative eigenvalue shifts remain below 0.021; retaining 75% of layers yields profile cosine similarities above 0.91 and JS divergences below 0.06. Broad cross-modal patterns persist as retained SVD energy varies from $E = 0.80$ to $E = 0.99$, although absolute spectral estimates depend on rank (Appendix C; Appendices C.2, C.4, C.3, and D).

Residual identity control. Residual connections can induce near-unit eigenvalues through identity propagation. To examine this effect, we apply TIMs to residual-update trajectories $\Delta_\ell = x_{\ell+1} - x_\ell$. Removing the explicit identity component shifts the absolute spectral distributions, confirming its contribution to near-unit concentration. However, raw-state and residual-update statistics remain correlated across models (Spearman $\rho = 0.584$, 0.619, and 0.774 for mean magnitude, median magnitude, and spectral radius; all $p \leq 0.0017$; Appendix E). Thus, residual propagation affects absolute spectral location but does not fully explain the relative organization across models.

5 Discussion

Cross-layer persistence reflects both general properties of representation propagation and components relevant to subsequent computation. Similar coarse spectral profiles on natural and Gaussian-random inputs suggest that the observed organization is not determined solely by stimulus semantics. At the same time, persistent representatives exhibit stronger terminal-output alignment, and their removal produces larger task-loss increases under the tested interventions. These results indicate that spectral persistence can identify functionally relevant directions without assuming that all persistent modes encode task-specific information. The cortical encoding results provide additional evidence of modality-dependent representational correspondence, particularly in audio and language.

Limitations. The affine surrogate approximates nonlinear, depth-varying trajectories, so $|\lambda|$ is an operational persistence score rather than an intrinsic property of the full network. Residual propagation contributes to near-unit eigenvalues, although it does not fully explain the observed cross-model organization. The analysis does not isolate the semantic content of individual modes or explain how persistence emerges during training. Cortical alignment does not establish shared neural mechanisms. Projection ablations demonstrate functional consequences under the tested interventions, but not universal necessity, and are limited to nine models, one intervention layer, and finite calibration and evaluation samples.

Future Work. Future work could extend interventions across layers and model families, including amplification and component replacement. Tracking TIMs during pretraining and fine-tuning may clarify how persistence emerges and relates to transfer and generalization. Extending the surrogate to nonlinear operators could further examine persistence across architectures and learning objectives.

6 Conclusion

We investigated whether deep neural networks exhibit a systematic organization of cross-layer persistence and introduced TIMs to quantify this previously undercharacterized property through trajectory-based spectral analysis. Across 108 pretrained models spanning vision, audio, and language, we identified systematic, modality-dependent persistence patterns that remain robust across experimental settings. Persistent representatives exhibit stronger terminal-output alignment, while removing persistent directions produces greater task-loss increases under projection ablation. Cortical alignment further reveals modality-dependent correspondence with biological responses, particularly in audio and language. Together, these findings establish cross-layer persistence as a measurable and systematically organized property of deep network representations and demonstrate its relevance to subsequent computation.

AI use statement

In this work, we used generative AI tools for manuscript language polishing and assisting with bug identification in research code. All research code was written by the authors, who independently investigated and resolved the identified issues. AI-assisted manuscript revisions were reviewed by the authors, and all code was tested for correctness. We did not use generative AI tools to generate research results or experimental data. We take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Reproducibility statement

We provide detailed descriptions of the TIMs framework, experimental settings, datasets, and evaluation protocols in the main text and appendices. Additional implementation details, robustness analyses, and per-model results are included in the appendices to facilitate reproducibility. All experiments are reproducible using the documented methodology and publicly available datasets and pretrained models. The source code and implementation will be made publicly available upon acceptance.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460, 2020.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens, 2023.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models With dictionary learning. Transformer Circuits Thread, 2023. URL `https://transformer-circuits.pub/2023/monosemantic-features`.

Bingni W. Brunton, Lise A. Johnson, Jeffrey G. Ojemann, and J. Nathan Kutz. Extracting spatial–temporal coherent patterns in large-scale neural recordings using dynamic mode decomposition. *Journal of Neuroscience Methods*, 258:1–15, 2016.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007.

James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

Akshunna S. Dogra and William T. Redman. Optimizing neural networks via koopman operator theory. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2087–2097, 2020.

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.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Vladimir Fonov, Alan C Evans, Kelly Botteron, C Robert Almli, Robert C McKinstry, and D Louis Collins. Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, 54(1): 313–327, 2011.

Karl J. Friston, Paul Fletcher, Oliver Josephs, Andrew Holmes, Michael D. Rugg, and Robert Turner. Event-related fMRI: Characterizing differential responses. *NeuroImage*, 7(1):30–40, 1998. doi: 10.1006/nimg.1997.0306.

Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. Technical report, California Institute of Technology, 2007.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse problems*, 34 (1):014004, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Martin N. Hebart, Oliver Contier, Lina Teichmann, Adam H. Rockter, Charles Y. Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I. Baker. THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12:e82580, 2023. doi: 10.7554/eLife.82580.

Martin N. Hebart, Oliver Contier, Lina Teichmann, Adam H. Rockter, Charles Y. Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I. Baker. THINGS-fMRI. OpenNeuro, 2024. URL https://openneuro.org/datasets/ds004192/versions/1.0.7.

Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature reviews neuroscience*, 8(5):393–402, 2007.

Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from ImageNet. GitHub repository, March 2019. URL https://github.com/fastai/imagenette.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Jean-Rémi King, Niccolò Pescetelli, and Stanislas Dehaene. Brain mechanisms underlying the brief maintenance of seen and unseen sensory information. *Neuron*, 92(5):1122–1134, 2016. doi: 10.1016/j.neuron.2016.10.051.

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.

Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G. Huth. A natural language fmri dataset for voxelwise encoding models. *Scientific Data*, 10(1):555, 2023. doi: 10.1038/s41597-023-02437-z.

Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G. Huth. An fmri dataset during a passive natural language listening task. OpenNeuro, 2025. URL https://openneuro.org/datasets/ds003020/versions/3.1.0.

Yann LeCun. A path towards autonomous machine intelligence. OpenReview, 2022. URL https://openreview.net/forum?id=BZ5a1r-kVsf. Version 0.9.2.

Jixing Li, Shohini Bhattasali, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, R. Nathan Spreng, Jonathan R. Brennan, Yiming Yang, Christophe Pallier, and John Hale. Le petit prince multilingual naturalistic fmri corpus. *Scientific Data*, 9(1):530, 2022. doi: 10.1038/s41597-022-01625-7.

Jixing Li, John Hale, and Christophe Pallier. Le petit prince: A multilingual fmri corpus using ecological stimuli. OpenNeuro, 2025. URL https://openneuro.org/datasets/ds003643/versions/2.0.7.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Igor Mezić. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, 41(1):309–325, 2005.

nostalgebraist. interpreting GPT: the logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in/.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Karol J. Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 1015–1018, 2015. doi: 10.1145/2733373.2806390.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and Boon Thau Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, 28(9):3095–3114, 2018.

Peter J. Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656:5–28, 2010.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

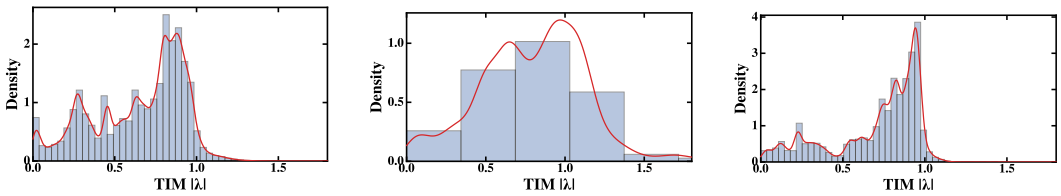
Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017.

Saurabh Vyas, Matthew D. Golub, David Sussillo, and Krishna V. Shenoy. Computation through neural population dynamics. *Annual Review of Neuroscience*, 43:249–275, 2020.

A TIM $|\lambda|$ ROBUSTNESS ACROSS INPUTS

The main text shows that persistent TIMs are prevalent across modalities (Section 4.2). A key interpretive question is whether these TIM $|\lambda|$ profiles are driven primarily by the semantic content of the input stimuli or reflect more general properties of the model's forward-processing pipeline. We address this with two controls: repeating the analysis on fMRI benchmark stimuli drawn from different natural stimulus sources, and evaluating Gaussian-random inputs that remove structured stimulus content.

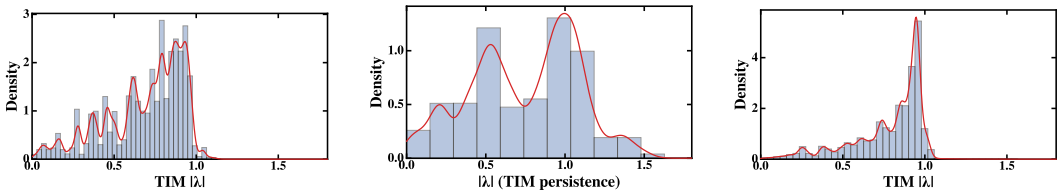
The fMRI benchmark stimuli produce TIM $|\lambda|$ distributions consistent with the standalone-dataset profiles in Figure 1, indicating that the observed modality-level persistence patterns are reproducible across natural stimulus sources.



(a) Audio (b) Vision (c) Language

Figure 5: TIM $|\lambda|$ distributions on the fMRI benchmark stimuli. The profiles are consistent with the standalone-dataset results in Figure 1.

We further evaluate Gaussian-random inputs as a synthetic-input control. These inputs reduce structured stimulus content while preserving each model's forward-processing pipeline. As shown in Figure 6 and Table 5, the KDE mode and mean $|\lambda|$ are not significantly different from the standalone-stimulus results across modalities.



(a) Audio (b) Vision (c) Language

Figure 6: TIM $|\lambda|$ distributions under Gaussian-random inputs. We randomly sample 10,000 synthetic inputs per modality.

Table 5: Standalone-versus-random comparison of TIM $|\lambda|$ summaries. KDE mode and mean $|\lambda|$ show no significant differences across modalities.

Modality	**Mode**		**Mean $\lvert\lambda\rvert$**		p_{mode}	p_{mean}
	Stand.	**Rand.**	**Stand.**	**Rand.**		
Audio	0.809	0.784	0.678	0.684	0.593	0.774
Vision	0.957	0.893	0.978	0.936	0.395	0.562
Language	0.906	0.905	0.776	0.783	0.532	0.640

Together, these results suggest that coarse TIM $|\lambda|$ profiles are stable across natural stimulus sources and random-input controls. This indicates that the persistence structure identified in Section 4.2 is not driven solely by stimulus semantics. Rather, coarse TIM $|\lambda|$ profiles reflect architecture- and parameter-level regularities in how trained networks process information across depth. Stimulus-specific content may determine *which* information occupies each persistence setting, while the broad persistence profile reflects properties of the model's forward-processing pipeline.

Table 6: Detailed model-level comparison of transient and persistent mode prevalence. Δ denotes transient minus persistent proportion; p-values are from paired tests across models.

Modality	**Transient**	**Persistent**	Δ	**Cohen's** d	p**-value**
Audio	2.2%	17.8%	-15.6%	-0.85	< 0.0001
Vision	3.3%	24.7%	-21.4%	-0.71	< 0.0001
Language	1.0%	42.5%	-41.5%	-1.77	< 0.0001

B MODEL-LEVEL TIM λ DISTRIBUTIONS

Figure 7 reports model-level TIM λ distributions for 18 representative models across audio, vision, and language.

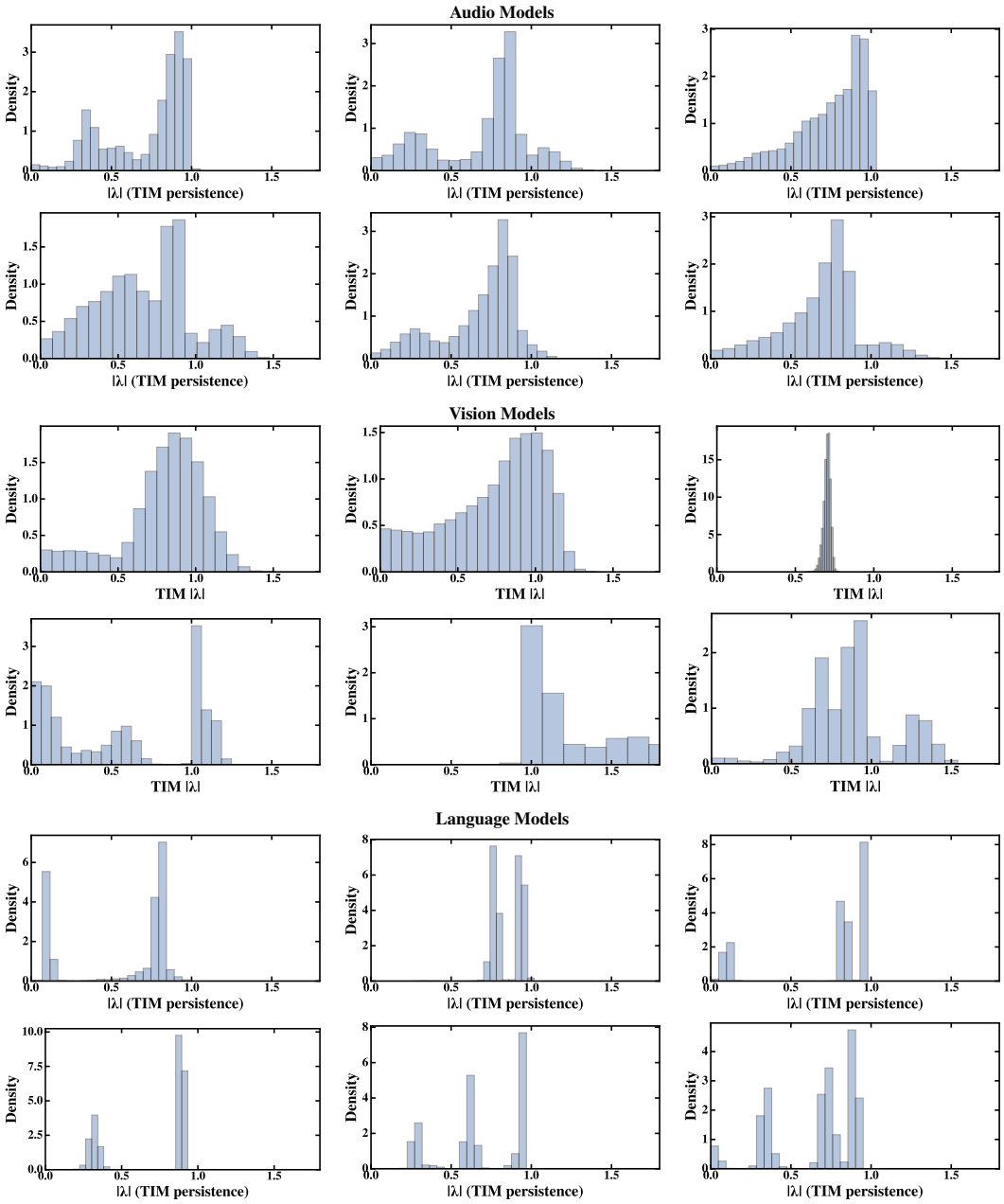


Figure 7: Model-level TIM λ distributions for a subset of 18 representative models, complementing the aggregate modality-level profiles in Fig. 1.

C Robustness Analysis of TIM Extraction

C.1 Held-Out Affine Transition Prediction

Table 7: Held-out transition prediction of the trajectory-restricted affine surrogate across modalities. Every third transition pair is excluded from operator estimation. ΔR^2 is computed relative to a layer-shuffle baseline. Proportions report 95% Clopper–Pearson confidence intervals; p-values use paired one-sided Wilcoxon tests.

	R^2 Statistics			E_{rel} Statistics		
Modality	**Mean ± Std**	ΔR^2	$R^2{>}0.90$	**Mean ± Std**	$E_{\mathrm{rel}}{<}0.10$	p**-value**
Audio	0.918 ± 0.236	$+0.910$	69.9% $[66.3, 73.4]$	0.128 ± 0.089	46.8% $[43.0, 50.6]$	1.2×10^{-113}
Vision	0.893 ± 0.169	$+0.869$	73.0% $[68.6, 77.2]$	0.147 ± 0.132	42.4% $[37.7, 47.2]$	3.8×10^{-73}
Language	0.931 ± 0.033	$+0.780$	85.6% $[82.4, 88.5]$	0.137 ± 0.065	31.2% $[27.3, 35.3]$	8.4×10^{-90}

These statistics quantify prediction quality on transition pairs excluded from operator estimation. Because endpoint states may occur in adjacent fitted transition pairs, they measure generalization to excluded transitions rather than to entirely unseen states.

C.2 Perturbation Stability

We first test whether TIM $|\lambda|$ estimates are stable under small additive perturbations to the snapshot matrices. For each modality, we vary the perturbation magnitude ϵ, re-extract TIMs, and report the average relative shift in mean $|\lambda|$; lower values indicate stronger numerical stability.

Table 8: Perturbation stability of TIM $|\lambda|$ estimates across modalities. Values report the average relative shift in mean $|\lambda|$ under different perturbation magnitudes ϵ.

Modality	$\epsilon = 0.001$	$\epsilon = 0.01$	$\epsilon = 0.05$
Audio	0.0001	0.0014	0.0115
Vision	0.0002	0.0020	0.0202
Language	0.0001	0.0012	0.0131

Across modalities, the relative shift in mean $|\lambda|$ remains small as ϵ increases. Even at $\epsilon = 0.05$, the average shift stays below 0.021, suggesting that TIM $|\lambda|$ estimates are not dominated by numerical fluctuations.

C.3 Layer-Sampling Sensitivity

We further assess whether the estimated persistence profiles depend on densely sampling network depth. We retain 75% of the original layers, recompute the modality-level profiles, and compare them with the full-layer estimates using cosine similarity and Jensen–Shannon (JS) divergence. We also compare model-level mean-$|\lambda|$ rankings using Spearman correlation.

Table 9: Stability under 75% layer sampling. Higher profile cosine similarity, lower JS divergence, and higher model-level Spearman correlation indicate closer agreement with the full-layer results.

Modality	**Profile cosine**	**JS divergence**	**Model-level Spearman** ρ
Audio	0.943	0.029	0.933
Language	0.913	0.053	0.701
Vision	0.949	0.022	0.972

With moderate layer subsampling, the persistence profiles remain close to the full-layer results: profile cosine similarities exceed 0.91, JS divergences remain below 0.06, and model-level rankings remain positively associated across all three modalities. The agreement is strongest for audio and vision and weaker, though still positive, for language. These results support robustness to moderate

layer subsampling; more aggressive subsampling may reduce fidelity and is not assumed to preserve the full-layer estimates.

C.4 Per-Model Perturbation Results

Tables 10–12 report per-model perturbation results for audio, vision, and language models. The low average sensitivity in Table 8 is broadly reflected across individual models. Values report the relative shift in mean $|\lambda|$ under each perturbation magnitude.

Table 10: Per-model perturbation stability of TIM $|\lambda|$ estimates for **audio** models. Values report the relative shift in mean $|\lambda|$ under each perturbation magnitude.

Model	$\epsilon = 0.001$	$\epsilon = 0.01$	$\epsilon = 0.05$
data2vec-audio-base	0.0001	0.0009	0.0064
data2vec-audio-base-960h	0.0002	0.0015	0.0094
data2vec-audio-large	0.0001	0.0010	0.0079
data2vec-audio-large-960h	0.0001	0.0016	0.0229
hubert-base-ls960	0.0001	0.0016	0.0079
hubert-base-superb-ks	0.0002	0.0015	0.0083
hubert-large-ls960-ft	0.0001	0.0005	0.0026
hubert-xlarge-ls960-ft	0.0000	0.0003	0.0146
sew-d-mid-100k	0.0001	0.0009	0.0049
sew-d-small-100k	0.0001	0.0006	0.0033
sew-d-tiny-100k	0.0001	0.0016	0.0084
sew-mid-100k	0.0002	0.0020	0.0102
sew-small-100k	0.0002	0.0020	0.0103
sew-tiny-100k	0.0001	0.0015	0.0082
unispeech-large-1500h-cv	0.0000	0.0008	0.0068
unispeech-sat-base	0.0002	0.0015	0.0088
unispeech-sat-base-plus	0.0001	0.0013	0.0082
unispeech-sat-large	0.0001	0.0014	0.0067
w2v-bert-2.0	0.0002	0.0015	0.0082
wav2vec2-base	0.0002	0.0022	0.0117
wav2vec2-base-960h	0.0006	0.0034	0.0133
wav2vec2-base-superb-ks	0.0002	0.0020	0.0103
wav2vec2-conformer-rel-pos-large	0.0001	0.0018	0.0117
wav2vec2-conformer-rope-large-960h-ft	0.0001	0.0009	0.0059
wav2vec2-large	0.0001	0.0012	0.0065
wav2vec2-large-960h	0.0001	0.0009	0.0057
wav2vec2-large-xlsr-53	0.0000	0.0001	0.0006
wav2vec2-xls-r-1b	0.0000	0.0004	0.0026
wav2vec2-xls-r-300m	0.0001	0.0030	0.0771
wavlm-base	0.0001	0.0014	0.0080
wavlm-base-plus	0.0001	0.0005	0.0030
wavlm-large	0.0001	0.0008	0.0065
whisper-base	0.0004	0.0039	0.0192
whisper-medium	0.0000	0.0001	0.0004
whisper-small	0.0000	0.0013	0.0516
whisper-tiny	0.0003	0.0031	0.0163

Table 11: Per-model perturbation stability of TIM $|\lambda|$ estimates for **vision** models. Values report the relative shift in mean $|\lambda|$ under each perturbation magnitude.

Model	$\epsilon = 0.001$	$\epsilon = 0.01$	$\epsilon = 0.05$
beit-base-patch16-224-pt22k-ft22k	0.0000	0.0012	0.0169
beit-large-patch16-224-pt22k-ft22k	0.0000	0.0003	0.0046
clip-vit-base-patch16	0.0001	0.0013	0.0158
clip-vit-base-patch32	0.0001	0.0016	0.0157
clip-vit-large-patch14	0.0001	0.0012	0.0124
convnext_base	0.0001	0.0010	0.0051
convnext_tiny	0.0001	0.0012	0.0059
data2vec-vision-base	0.0000	0.0003	0.0030
data2vec-vision-large	0.0000	0.0000	0.0002
deit-base-patch16-224	0.0001	0.0007	0.0102
deit-small-patch16-224	0.0001	0.0011	0.0065
densenet121	0.0001	0.0007	0.0037
densenet201	0.0001	0.0008	0.0041
dino-vitb16	0.0001	0.0007	0.0039
dino-vits16	0.0000	0.0004	0.0019
dinov2-base	0.0003	0.0032	0.0186
dinov2-large	0.0005	0.0031	0.0387
dinov2-small	0.0004	0.0047	0.0462
efficientnet_b0	0.0007	0.0073	0.0359
efficientnet_b4	0.0004	0.0041	0.0203
resnet50	0.0001	0.0007	0.0037
resnet101	0.0001	0.0006	0.0031
sam-vit-base	0.0001	0.0011	0.0124
sam-vit-large	0.0000	0.0001	0.0101
sam-vit-huge	0.0000	0.0001	0.0039
swin-tiny-patch4-window7-224	0.0009	0.0092	0.1308
swin-small-patch4-window7-224	0.0003	0.0050	0.0557
swin-large-patch4-window7-224	0.0004	0.0030	0.0434
vgg16	0.0001	0.0008	0.0040
vgg19	0.0001	0.0010	0.0052
vit-base-patch16-224-in21k	0.0000	0.0003	0.0089
vit-large-patch16-224-in21k	0.0000	0.0005	0.0064
vit-mae-base	0.0000	0.0005	0.0042
vit-mae-large	0.0000	0.0006	0.0293
vit-msn-base	0.0002	0.0020	0.0188
vit-msn-large	0.0005	0.0102	0.1172

Table 12: Per-model perturbation stability of TIM $|\lambda|$ estimates for **language** models. Values report the relative shift in mean $|\lambda|$ under each perturbation magnitude.

Model	$\epsilon = 0.001$	$\epsilon = 0.01$	$\epsilon = 0.05$
MiniLM-L6-H384-uncased	0.0005	0.0043	0.0458
all-MiniLM-L6-v2	0.0001	0.0011	0.0096
all-mpnet-base-v2	0.0001	0.0008	0.0052
albert-base-v2	0.0001	0.0007	0.0049
albert-large-v2	0.0000	0.0008	0.0095
albert-xlarge-v2	0.0001	0.0008	0.0110
bert-base-cased	0.0002	0.0024	0.0135
bert-base-multilingual-cased	0.0000	0.0004	0.0044
bert-base-uncased	0.0000	0.0006	0.0031
bert-large-cased	0.0000	0.0003	0.0050
bert-large-uncased	0.0001	0.0008	0.0057
camembert-base	0.0002	0.0015	0.0102
conv-bert-base	0.0002	0.0018	0.0108
conv-bert-medium-small	0.0004	0.0024	0.0152
data2vec-text-base	0.0001	0.0007	0.0071
deberta-base	0.0001	0.0017	0.0152
deberta-large	0.0000	0.0003	0.0072
distilbert-base-multilingual-cased	0.0000	0.0004	0.0019
distilbert-base-uncased	0.0000	0.0005	0.0029
distilroberta-base	0.0001	0.0010	0.0089
electra-base-discriminator	0.0001	0.0009	0.0048
electra-large-discriminator	0.0000	0.0003	0.0043
electra-small-discriminator	0.0001	0.0012	0.0085
ernie-2.0-base-en	0.0001	0.0008	0.0060
ernie-2.0-large-en	0.0000	0.0002	0.0040
ibert-roberta-base	0.0001	0.0029	0.0536
mpnet-base	0.0001	0.0007	0.0055
rembert	0.0001	0.0008	0.0096
roberta-base	0.0001	0.0029	0.0536
roberta-large	0.0000	0.0004	0.0080
squeezebert-uncased	0.0002	0.0038	0.0361
t5-small	0.0001	0.0010	0.0074
xlm-roberta-base	0.0001	0.0006	0.0093
xlm-roberta-large	0.0000	0.0005	0.0105
xlnet-base-cased	0.0001	0.0005	0.0325
xlnet-large-cased	0.0001	0.0018	0.0219

D RANK-SELECTION SENSITIVITY

We next test whether modality-level TIM $|\lambda|$ profiles depend on a single SVD rank-selection threshold. We vary the retained-energy threshold from $E = 0.80$ to $E = 0.99$ and report the resulting mean TIM $|\lambda|$ across models within each modality.

Table 13: Sensitivity of TIM $|\lambda|$ estimates to energy-based rank selection. Values report mean TIM $|\lambda|$ across models, with standard deviations in parentheses.

Modality	$E = 0.80$	$E = 0.85$	$E = 0.90$	$E = 0.95$	$E = 0.99$
Audio	0.637 ± 0.151	0.649 ± 0.148	0.673 ± 0.133	0.731 ± 0.111	0.824 ± 0.103
Vision	1.042 ± 0.725	1.014 ± 0.652	0.979 ± 0.623	0.921 ± 0.452	0.889 ± 0.343
Language	0.773 ± 0.182	0.757 ± 0.141	0.775 ± 0.086	0.811 ± 0.047	0.888 ± 0.063

The absolute mean TIM $|\lambda|$ varies with the retained-energy threshold, especially for vision at lower thresholds and for audio at higher thresholds. However, the qualitative modality-level pattern remains stable across thresholds, indicating that the TIM $|\lambda|$ conclusions are not driven by a single energy-based rank-selection setting.

E RESIDUAL-UPDATE SPECTRAL CONTROL

To quantify how much residual identity propagation contributes to the raw-state spectra, we apply TIMs to the residual-update trajectories $\Delta_\ell = x_{\ell+1} - x_\ell$ and compare matched model-level statistics. Removing the explicit identity component shifts the absolute spectra, while several relative model-level properties remain positively associated with the raw-state analysis.

Table 14: Association between model-level spectral statistics computed from raw hidden-state trajectories and residual-update trajectories.

Spectral statistic	**Spearman ρ**	**p-value**
Magnitude closest to one	0.392	0.0476
Mean magnitude	0.584	0.0017
Median magnitude	0.619	0.0007
Spectral radius	0.774	< 0.0001

F Complementary In-Trajectory Retention Analysis

As a complementary analysis to the stricter terminal-held-out evaluation in Section 4.3, we construct TIM representatives from the complete observed trajectory and measure cosine similarity to the corresponding model-output representation. Differences across persistence settings are tested at the model level using Friedman tests and post-hoc paired Wilcoxon signed-rank tests with Cohen's d.

Table 15: In-trajectory comparison of model-output alignment across TIM persistence settings.

Modality	#Models	Friedman $\chi^2(2)$	Test	Mean Δ	Cohen's d	W-stat	p-value	Sig
Audio	36	13.92 ($p < 0.001^{***}$)	P > T	+0.1521	0.75	105.0	< 0.001	***
			P > M	+0.0805	0.33	219.0	0.0743	n.s.
			M > T	+0.0716	0.54	98.0	< 0.001	***
Vision	36	58.16 ($p < 0.0001^{***}$)	P > T	+0.2615	1.22	0.0	< 0.0001	***
			P > M	+0.2229	1.02	13.0	< 0.0001	***
			M > T	+0.0387	0.47	18.0	< 0.01	**
Language	36	31.99 ($p < 0.0001^{***}$)	P > T	+0.2068	1.20	21.0	< 0.0001	***
			P > M	+0.1051	0.47	178.0	< 0.05	*
			M > T	+0.1018	0.76	40.0	< 0.0001	***

Persistent-setting representatives show higher aggregate output alignment than transient-setting representatives in all three modalities. The weaker P > M separation in audio is consistent with the broader audio persistence profile. Because the complete trajectory includes the terminal state, we treat this analysis as complementary evidence and rely on the terminal-held-out experiment in the main text for the primary retention claim.

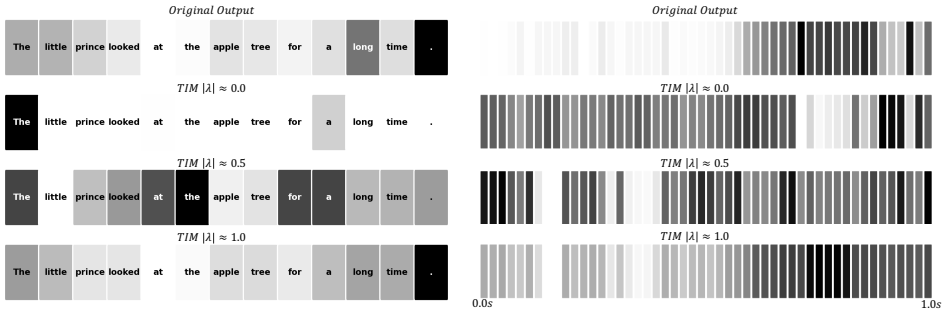


Figure 8: Representative PCA visualization of TIM reconstructions from BERT-Base (Devlin et al., 2019) and Wav2Vec2-Base (Baevski et al., 2020) across persistence settings.

Figure 8 complements the vision example in the main text by illustrating the reconstructed geometry for representative language and audio models.

G PER-MODEL IN-TRAJECTORY RETENTION RESULTS

This section reports the per-model retention scores underlying the modality-level statistics in Table 15. For each model, we report the mean cosine similarity between TIM-reconstructed representations and the corresponding model-output representations under three target persistence settings: transient (T, $\gamma = 0.0$), mid-persistence (M, $\gamma = 0.5$), and persistent (P, $\gamma = 1.0$). The $\Delta(P{-}T)$ column measures the improvement of the persistent setting over the transient setting, and the Friedman test evaluates whether retention differs across the three settings within each model.

Table 16: Per-model model-output retention for **audio** models under different target persistence settings.

Model	T ($\gamma = 0.0$)	M ($\gamma = 0.5$)	P ($\gamma = 1.0$)	$\Delta(P{-}T)$	χ^2	p
data2vec-audio-base	0.0019	0.2605	0.0961	0.0942	2676.3	$< 0.0001^{***}$
data2vec-audio-base-960h	0.0163	0.3379	0.4115	0.3952	3337.1	$< 0.0001^{***}$
data2vec-audio-large	0.0386	0.1198	0.3044	0.2658	3152.6	$< 0.0001^{***}$
data2vec-audio-large-960h	0.0151	0.3294	0.0145	−0.0006	4051.8	$< 0.0001^{***}$
hubert-base-ls960	0.0000	0.0008	0.5561	0.5561	6997.6	$< 0.0001^{***}$
hubert-base-superb-ks	0.0000	0.0009	0.5632	0.5632	7073.6	$< 0.0001^{***}$
hubert-large-ls960-ft	0.1876	0.0632	0.0002	−0.1874	4689.7	$< 0.0001^{***}$
hubert-xlarge-ls960-ft	0.0598	0.0455	0.0119	−0.0479	499.5	$< 0.0001^{***}$
sew-d-mid-100k	0.0000	0.0003	0.1541	0.1541	4924.7	$< 0.0001^{***}$
sew-d-small-100k	0.0006	0.1208	0.3092	0.3086	3358.2	$< 0.0001^{***}$
sew-d-tiny-100k	0.0000	0.1004	0.1234	0.1234	1253.5	$< 0.0001^{***}$
sew-mid-100k	0.0000	0.0000	0.2346	0.2346	4852.0	$< 0.0001^{***}$
sew-small-100k	0.0030	0.2769	0.1192	0.1162	3466.7	$< 0.0001^{***}$
sew-tiny-100k	0.0000	0.1333	0.0199	0.0199	1886.5	$< 0.0001^{***}$
unispeech-large-1500h-cv	0.0000	0.1699	0.2672	0.2672	3964.4	$< 0.0001^{***}$
unispeech-sat-base	0.0000	0.0387	0.3361	0.3361	3532.8	$< 0.0001^{***}$
unispeech-sat-base-plus	0.0000	0.0281	0.4044	0.4044	4451.5	$< 0.0001^{***}$
unispeech-sat-large	0.0137	0.0915	0.0055	−0.0082	2238.3	$< 0.0001^{***}$
w2v-bert-2.0	0.0034	0.0000	0.0255	0.0221	1757.0	$< 0.0001^{***}$
wav2vec2-base	0.0003	0.0820	0.2034	0.2031	3065.2	$< 0.0001^{***}$
wav2vec2-base-960h	0.0450	0.0012	0.0009	−0.0441	476.0	$< 0.0001^{***}$
wav2vec2-base-superb-ks	0.0005	0.0818	0.2158	0.2153	3339.0	$< 0.0001^{***}$
wav2vec2-conformer-rel-pos-large	0.0315	0.0002	0.0955	0.0640	7812.0	$< 0.0001^{***}$
wav2vec2-conformer-rope-large-960h-ft	0.0000	0.0001	0.1024	0.1024	9792.0	$< 0.0001^{***}$
wav2vec2-large	0.0000	0.0000	0.2977	0.2977	9943.3	$< 0.0001^{***}$
wav2vec2-large-960h	0.0000	0.0000	0.3358	0.3358	9956.0	$< 0.0001^{***}$
wav2vec2-large-xlsr-53	0.1480	0.1821	0.0063	−0.1417	1089.6	$< 0.0001^{***}$
wav2vec2-xls-r-1b	0.0000	0.0008	0.5080	0.5080	9117.4	$< 0.0001^{***}$
wav2vec2-xls-r-300m	0.0054	0.1063	0.0034	−0.0020	4321.3	$< 0.0001^{***}$
wavlm-base	0.0000	0.0250	0.4267	0.4267	4616.3	$< 0.0001^{***}$
wavlm-base-plus	0.0005	0.2684	0.0308	0.0303	4046.6	$< 0.0001^{***}$
wavlm-large	0.0242	0.0726	0.0009	−0.0233	1317.9	$< 0.0001^{***}$
whisper-base	0.2250	0.2516	0.0107	−0.2143	10925.8	$< 0.0001^{***}$
whisper-medium	0.0012	0.0092	0.1481	0.1469	40459.0	$< 0.0001^{***}$
whisper-small	0.3649	0.1214	0.3465	−0.0184	7066.7	$< 0.0001^{***}$
whisper-tiny	0.0606	0.5039	0.0336	−0.0270	24299.5	$< 0.0001^{***}$

Table 17: Per-model model-output retention for **vision** models under different target persistence settings.

Model	T ($\gamma = 0.0$)	M ($\gamma = 0.5$)	P ($\gamma = 1.0$)	$\Delta(P{-}T)$	χ^2	p
beit-base-patch16-224-pt22k-ft22k	0.0479	0.1175	0.4298	0.3819	5865.9	$< 0.0001^{***}$
beit-large-patch16-224-pt22k-ft22k	0.0000	0.0089	0.6723	0.6723	18301.1	$< 0.0001^{***}$
clip-vit-base-patch16	0.0029	0.3239	0.4523	0.4494	6867.5	$< 0.0001^{***}$
clip-vit-base-patch32	0.0000	0.2349	0.5168	0.5168	2244.3	$< 0.0001^{***}$
clip-vit-large-patch14	0.0000	0.0062	0.7050	0.7050	25131.0	$< 0.0001^{***}$
data2vec-vision-base	0.0000	0.0001	0.2442	0.2442	17485.3	$< 0.0001^{***}$
data2vec-vision-large	0.0000	0.0000	0.4609	0.4609	19500.0	$< 0.0001^{***}$
deit-base-patch16-224	0.0001	0.0114	0.7358	0.7357	14916.0	$< 0.0001^{***}$
deit-small-patch16-224	0.0030	0.1323	0.1606	0.1576	2076.9	$< 0.0001^{***}$
dino-vitb16	0.0006	0.0000	0.3499	0.3493	8314.0	$< 0.0001^{***}$
dino-vits16	0.1187	0.0010	0.4554	0.3367	9414.9	$< 0.0001^{***}$
dinov2-base	0.1365	0.2555	0.3357	0.1992	2757.5	$< 0.0001^{***}$
dinov2-large	0.0660	0.2438	0.3990	0.3330	4973.4	$< 0.0001^{***}$
dinov2-small	0.1298	0.2629	0.2718	0.1420	1872.6	$< 0.0001^{***}$
sam-vit-base	0.0331	0.0776	0.5258	0.4927	26655.9	$< 0.0001^{***}$
sam-vit-large	0.0000	0.0000	0.1291	0.1291	68398.0	$< 0.0001^{***}$
sam-vit-huge	0.0000	0.0000	0.2582	0.2582	99970.0	$< 0.0001^{***}$
vit-base-patch16-224-in21k	0.0086	0.0630	0.3912	0.3826	7117.3	$< 0.0001^{***}$
vit-large-patch16-224-in21k	0.0062	0.0000	0.2672	0.2610	19415.7	$< 0.0001^{***}$
vit-mae-base	0.0023	0.1665	0.0980	0.0957	1145.0	$< 0.0001^{***}$
vit-mae-large	0.0000	0.0000	0.3007	0.3007	4762.0	$< 0.0001^{***}$
vit-msn-base	0.0004	0.0128	0.7197	0.7193	15107.1	$< 0.0001^{***}$
vit-msn-large	0.0995	0.1292	0.1274	0.0279	9249.4	$< 0.0001^{***}$
resnet18	0.0000	0.0000	0.0868	0.0868	4900.0	$< 0.0001^{***}$
resnet34	0.0000	0.0000	0.0873	0.0873	4900.0	$< 0.0001^{***}$
resnet50	0.0000	0.0000	0.0453	0.0453	4892.0	$< 0.0001^{***}$
resnet101	0.0000	0.0000	0.0496	0.0496	4882.0	$< 0.0001^{***}$
wide_resnet50_2	0.0000	0.0000	0.0479	0.0479	4894.0	$< 0.0001^{***}$
densenet121	0.0000	0.0000	0.0465	0.0465	4900.0	$< 0.0001^{***}$
densenet201	0.0000	0.0000	0.0452	0.0452	4900.0	$< 0.0001^{***}$
efficientnet_b0	0.0000	0.0000	0.1604	0.1604	4900.0	$< 0.0001^{***}$
efficientnet_b4	0.0000	0.0000	0.1530	0.1530	4900.0	$< 0.0001^{***}$
convnext_tiny	0.0000	0.0000	0.0617	0.0617	4900.0	$< 0.0001^{***}$
convnext_base	0.0000	0.0000	0.1491	0.1491	4900.0	$< 0.0001^{***}$
vgg16	0.0000	0.0000	0.0669	0.0669	4900.0	$< 0.0001^{***}$
vgg19	0.0000	0.0000	0.0645	0.0645	4900.0	$< 0.0001^{***}$

Table 18: Per-model model-output retention for **language** models under different target persistence settings.

Model	T ($\gamma = 0.0$)	M ($\gamma = 0.5$)	P ($\gamma = 1.0$)	$\Delta(P{-}T)$	χ^2	p
albert-base-v2	0.0109	0.1852	0.0122	0.0013	4781.0	$< 0.0001^{***}$
albert-large-v2	0.0011	0.1525	0.2258	0.2247	10454.8	$< 0.0001^{***}$
albert-xlarge-v2	0.0001	0.0670	0.4726	0.4725	15353.1	$< 0.0001^{***}$
MiniLM-L6-H384-uncased	0.0199	0.0099	0.0001	-0.0198	4773.6	$< 0.0001^{***}$
all-mpnet-base-v2	0.0000	0.1884	0.1010	0.1010	10288.2	$< 0.0001^{***}$
bert-base-cased	0.0000	0.1352	0.2517	0.2517	10928.5	$< 0.0001^{***}$
bert-base-multilingual-cased	0.0000	0.4540	0.0863	0.0863	16066.5	$< 0.0001^{***}$
bert-base-uncased	0.0000	0.1635	0.1440	0.1440	9788.8	$< 0.0001^{***}$
bert-large-cased	0.0000	0.0002	0.2501	0.2501	19959.3	$< 0.0001^{***}$
bert-large-uncased	0.0000	0.0000	0.4124	0.4124	20000.0	$< 0.0001^{***}$
camembert-base	0.0000	0.3072	0.1659	0.1659	14291.8	$< 0.0001^{***}$
conv-bert-base	0.0000	0.0203	0.2098	0.2098	11416.9	$< 0.0001^{***}$
conv-bert-medium-small	0.0000	0.0592	0.1226	0.1226	14153.0	$< 0.0001^{***}$
data2vec-text-base	0.0000	0.2642	0.1940	0.1940	6318.1	$< 0.0001^{***}$
deberta-base	0.0000	0.0177	0.0038	0.0038	556.3	$< 0.0001^{***}$
deberta-large	0.0000	0.0000	0.2446	0.2446	19706.0	$< 0.0001^{***}$
distilbert-base-multilingual-cased	0.0000	0.1000	0.0003	0.0003	13817.6	$< 0.0001^{***}$
distilbert-base-uncased	0.0402	0.0016	0.0407	0.0005	4117.7	$< 0.0001^{***}$
distilroberta-base	0.0000	0.2541	0.1802	0.1802	4996.1	$< 0.0001^{***}$
electra-base-discriminator	0.0000	0.2416	0.3476	0.3476	9553.9	$< 0.0001^{***}$
electra-large-discriminator	0.0000	0.0000	0.4203	0.4203	19942.0	$< 0.0001^{***}$
electra-small-discriminator	0.0000	0.0124	0.0814	0.0814	10776.6	$< 0.0001^{***}$
ernie-2.0-base-en	0.0000	0.3542	0.1367	0.1367	11613.4	$< 0.0001^{***}$
ernie-2.0-large-en	0.0000	0.0000	0.4701	0.4701	20000.0	$< 0.0001^{***}$
ibert-roberta-base	0.0011	0.0005	0.3855	0.3844	15894.8	$< 0.0001^{***}$
mpnet-base	0.0000	0.1181	0.1264	0.1264	8417.3	$< 0.0001^{***}$
rembert	0.0000	0.0000	0.3550	0.3550	20000.0	$< 0.0001^{***}$
roberta-base	0.0011	0.0005	0.3855	0.3844	15894.8	$< 0.0001^{***}$
roberta-large	0.0000	0.0230	0.2727	0.2727	14774.1	$< 0.0001^{***}$
squeezebert-uncased	0.0000	0.0843	0.0469	0.0469	8811.5	$< 0.0001^{***}$
t5-small	0.0000	0.0496	0.6075	0.6075	16903.8	$< 0.0001^{***}$
xlm-roberta-base	0.0006	0.3109	0.1034	0.1028	9726.9	$< 0.0001^{***}$
xlm-roberta-large	0.0000	0.1187	0.2873	0.2873	16590.2	$< 0.0001^{***}$
xlnet-base-cased	0.2991	0.1199	0.1494	-0.1497	1215.4	$< 0.0001^{***}$
xlnet-large-cased	0.0029	0.2591	0.4549	0.4520	7375.7	$< 0.0001^{***}$
all-MiniLM-L6-v2	0.0483	0.0157	0.1223	0.0740	6950.2	$< 0.0001^{***}$