PROBING SELECTIVITY AND TOLERANCE WITH OPTIMIZED DYNAMIC STIMULI IN A FOUNDATION MODEL OF THE MOUSE VISUAL CORTEX

Anonymous authors
Paper under double-blind review

ABSTRACT

Selectivity and tolerance are fundamental response properties of visual systems. Selectivity describes how concentrated responses are across a stimulus set, whereas tolerance describes response preservation under a specified stimulus transformation. Comparing these properties across visual areas with a common dynamic stimulus protocol remains difficult when reference stimuli are restricted to predefined stimulus families. We analyze the MICrONS-adapted instance of a pretrained video-response foundation model as an in silico hypothesis generator. For each of 104,171 model readouts, we synthesize a most exciting dynamic input (MEDI), a short video optimized in a generative latent space. Compared with static stimuli, MEDIs contain temporal structure, enabling temporal transformations such as reversal and speed resampling. The readout-by-MEDI matrix shows positive within-model-area enrichment, while MEDI-set stimulus selectivity is the highest in model primary visual cortex (V1) and the lowest in model rostrolateral area (RL), with model lateromedial area (LM) and model anterolateral area (AL) intermediate; this pattern is not a simple monotonic progression along the established V1→LM/RL→AL hierarchy. Transforming each readout's own MEDI reveals transformation-specific tolerance: model RL retains relatively more response under strong spatial transformations, whereas model V1 retains relatively more under slow-down and relatively less under strong speed-up. Across transformations, tolerance does not show a single consistent area ordering. Selectivity and tolerance are negatively associated at selected transformation strengths, but the magnitude of this association depends on transformation and parameter rather than defining a single universal axis. The results revealed profound properties of different populations of readouts of the foundation model and provided experimentally testable predictions for the corresponding visual areas.

1 INTRODUCTION

Selectivity and tolerance describe complementary aspects of visual responses: selectivity measures how concentrated a neuron's responses are across stimuli, whereas tolerance measures how well those responses are preserved under a specified transformation. Their relationship, however, has not been consistently characterized across studies. In macaque inferotemporal cortex (IT), stronger object selectivity was associated with lower tolerance to changes in position, size, contrast, and clutter (Zoccolan et al., 2007). Yet this trade-off does not appear to hold uniformly: more selective IT neurons can also preserve their relative stimulus preferences more strongly across changes in size, position, and orientation (Zhivago & Arun, 2016). Moreover, across macaque V4 and IT, sensitivity to feature conjunctions showed no significant association with tolerance to changes in position, size, and background (Rust & DiCarlo, 2012). Together, these findings leave the relationship between selectivity and tolerance unresolved.

One limitation of existing approaches is that the stimuli used to probe this relationship are typically defined in advance by the experimenter. Neural tuning itself can depend on the stimulus set used to probe a neuron: different stimulus sets can yield substantially different receptive-field estimates (Yeh et al., 2009), and the same V4 neurons can exhibit markedly different tuning profiles when probed with different stimulus types (Hegdé & Van Essen, 2005). This limitation is particularly important for

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.

tolerance, because tolerance is usually defined relative to a reference stimulus whose response is then tested under transformation. A predefined stimulus family may restrict which response-driving inputs can serve as references, and therefore may not reveal how a neuron preserves its response around an input that strongly drives it.

A response-predictive model can overcome this limitation by allowing the reference stimulus itself to be optimized rather than selected from a fixed stimulus family. Deep predictive models have been used to synthesize most exciting static inputs (MESIs), revealing response-driving structure beyond conventional Gabor-like parameterizations in mouse V1 (Walker et al., 2019). More recently, the MICrONS functional resource and its video-response foundation model have enabled prediction of neural responses to diverse dynamic visual stimuli (MICrONS Consortium, 2025; Wang et al., 2025). This makes it possible to optimize neuron-specific dynamic stimuli directly in a high-dimensional video space, enabling tolerance to dynamic transformations beyond the static transformations studied in much of the selectivity–tolerance literature (Zoccolan et al., 2007; Rust & DiCarlo, 2010; 2012; Sharpee et al., 2013; Ziemba et al., 2016).

We therefore applied this framework to the video-response foundation model (Wang et al., 2025). For each of 104,171 model readouts resembling neurons in primary visual cortex (V1), lateromedial area (LM), rostrolateral area (RL), or anterolateral area (AL), we synthesized a most exciting dynamic input (MEDI). A shared MEDI set provided a common basis for quantifying selectivity, while transformations of each readout's own MEDI quantified tolerance to contrast, rotation, translation, temporal reversal, speed-up, and slow-down. At selected transformation strengths, selectivity was negatively associated with tolerance across all six transformations, but the magnitude of this association varied across transformations and parameters rather than following a single universal relationship. These results provided testable predictions for the corresponding visual areas.

2 Background and Related Work

Selectivity and tolerance. Selectivity may denote feature preference, stimulus discrimination, or response concentration across a stimulus set. We use finite-set-corrected lifetime sparseness, a standard response-concentration metric (Rolls & Tovee, 1995; Vinje & Gallant, 2000; Olshausen & Field, 2004), that is low for broadly distributed responses and high when a small subset of stimuli accounts for more of the response. Tolerance asks how much of a response or representation is preserved under a specified change. A large body of work has characterized these properties using experimenter-defined stimulus families, including object images and spatial transformations in macaque IT (Zoccolan et al., 2007), curvature stimuli in V4 (Sharpee et al., 2013), texture samples and metamers (Ziemba et al., 2016; Freeman & Simoncelli, 2011), and shape or object stimuli (Okazawa et al., 2015; Froudarakis et al., 2026). Related mouse studies have likewise probed texture (Yu et al., 2022), spatial integration (Murgas et al., 2020), and spatiotemporal processing (Han et al., 2022) using predefined stimulus sets.

The central limitation is that these approaches probe neural responses within stimulus spaces defined a priori. Neural tuning can depend on the stimulus set or stimulus type used for probing: different stimulus sets can yield substantially different receptive-field estimates (Yeh et al., 2009), and the same V4 neurons can exhibit markedly different tuning profiles across stimulus types (Hegdé & Van Essen, 2005). For tolerance, this dependence extends to the choice of reference stimulus, since preservation is measured relative to the particular input being transformed. Thus, predefined stimulus families may constrain the response-driving inputs available for measuring tolerance, motivating approaches that can identify response-relevant stimuli beyond a fixed stimulus parameterization.

Optimized stimuli. The most exciting static input (MESI) is an image optimized to drive a selected neuron or model unit. Deep generative and predictive models now synthesize such response-driving inputs (Nguyen et al., 2016a;b; Walker et al., 2019; Ponce et al., 2019; Xiao & Kreiman, 2020; Pierzchlewicz et al., 2023; Fel et al., 2023). Recent work maps tuning landscapes and level sets (Wang & Ponce, 2022; 2023), invariance manifolds (Bashiri et al., 2025), and functional invariance in mouse V1 (Ding et al., 2026). These studies focus on the geometry or invariance structure of response landscapes, whereas we ask how response concentration relates to transformation-specific response retention across visual areas.

Model-based optimization has also been extended to dynamic stimuli. Höfling et al. optimized time-varying stimuli for individual retinal neurons using a response-predictive model, revealing nonlinear chromatic selectivity from naturalistic video inputs (Höfling et al., 2024); Burg et al. optimized dynamic most discriminative stimuli to distinguish response-defined retinal cell groups (Burg et al., 2024). NEvo further used neural-guided video optimization to characterize dynamic selectivity at the level of human-fMRI visual regions (Tang et al., 2026). These studies establish that response-predictive models can optimize dynamic stimuli, but they address different questions: stimulus optimization for individual retinal selectivity, cell-group discrimination, or regional dynamic selectivity. Here, we instead use optimized dynamic references to study the relationship between selectivity and transformation-specific tolerance across model visual areas.

The MICrONS dataset and the foundation model. The MICrONS resource provides large-scale functional recordings across mouse visual cortex under shared natural movies, with dense anatomical reconstruction for a subset of neurons (MICrONS Consortium, 2025). Its video-response foundation model predicts time-varying activity using a shared spatiotemporal representation, behavioral covariates, and readout-specific parameters (Wang et al., 2025). This common model coordinate system permits one optimization and transformation protocol across areas. We use it as an in silico hypothesis generator: MEDIs and all reported response properties are conditional on the fitted model.

Functional organization of the mouse visual hierarchy. Anatomical and connectivity analyses support a hierarchical organization of mouse visual cortex, with V1 at an early level, LM and RL at an intermediate level, and AL at a higher level (Harris et al., 2019; Siegle et al., 2021; D'Souza et al., 2022). Functional studies have likewise reported systematic differences across visual areas in spatial and temporal frequency preferences, motion and texture processing, spatial integration, and response dynamics (Marshel et al., 2011; Andermann et al., 2011; Yu et al., 2022; Murgas et al., 2020; Han et al., 2022). Studies along rodent visual pathways have also reported increasing position tolerance and temporal stability at some hierarchical transitions (Vermaercke et al., 2014; Piasini et al., 2021). LM and RL are treated here as a shared intermediate level rather than as successive stages, consistent with connectivity-based hierarchy estimates (D'Souza et al., 2022). We therefore use V1→LM/RL→AL as an external biological reference and ask whether MEDI-derived selectivity and tolerance show simple organization along this axis.

3 METHODS

3.1 MEDI GENERATION

We used the publicly released video-response foundation model trained on MICrONS natural-movie recordings (Wang et al., 2025; MICrONS Consortium, 2025). The model combined visual input, behavioral modulation, a spatiotemporal core, and readout-specific parameters to predict time-varying activity. Throughout, model V1, model LM, model RL, and model AL denoted groups of model readouts associated with neurons recorded in the corresponding visual areas. We treated the foundation model as a digital twin for hypothesis generation only.

Figure 1 summarizes the iterative optimization pipeline. We used TAESD (Tiny Autoencoder for Stable Diffusion) as the pretrained autoencoder (Bohan, 2023). The robustness of this latent-space choice was assessed in Appendix A.1. For each target readout, we optimized a sequence of $T = 60$ latent tensors $z_1, \ldots, z_T \in \mathbb{R}^{4\times36\times64}$ (channels $\times$ height $\times$ width) in the latent space of a pretrained autoencoder. The corresponding video frames were $x_1, \ldots, x_T \in \mathbb{R}^{144\times256}$, where each $x_t = D(z_t)$ was decoded by the autoencoder. The objective maximized the target readout's mean predicted response while spatial and temporal smoothness penalties regularized the decoded sequence. Optimization used overlapping 15-frame windows with 8-frame overlap: shared frames were copied from the preceding window, while newly exposed frames started from the common gray-background latent initialization. Independent latent noise was injected before each update, preserving temporal continuity while keeping the model's recurrent state and GPU memory bounded. We used SGD with momentum and cosine learning-rate decay; the exact loss, weights, and initialization are given in Appendix B.1. Pixel-space optimization produced conspicuous high-frequency artifacts, so all production MEDIs used the latent-space procedure.

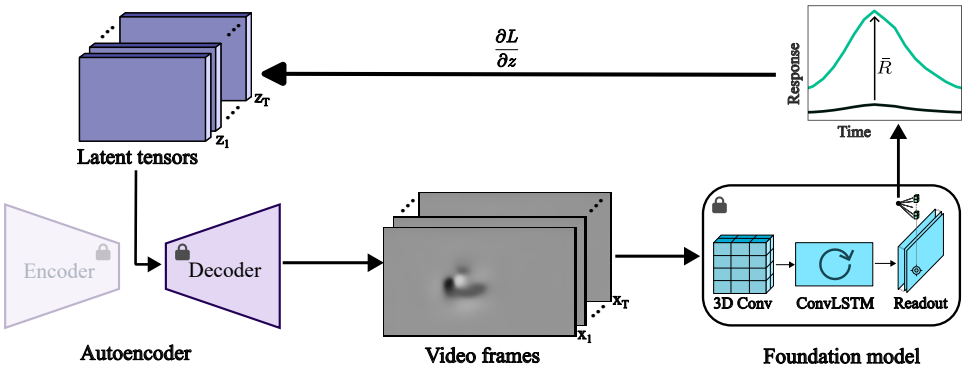


Figure 1: **Iterative optimization pipeline for generating MEDI.** Video frames are passed to the model, and the loss is backpropagated.

3.2 Dynamic Gabor stimulus generation

To validate that MEDIs elicited larger responses than a traditional low-dimensional Gabor stimulus, we generated a 60-frame drifting Gabor for each target readout. We jointly optimized its orientation, Gaussian width, wavelength, phase, aspect ratio, contrast, center position, temporal frequency, and image velocity against the same frozen foundation model used for MEDIs. Exact bounds, optimization, projection, and response extraction are given in Appendix B.2.

3.3 MEDI-set stimulus selectivity and transformation tolerance

For MEDI-set stimulus selectivity, we normalized a target readout's responses within each source-area MEDI pool by that pool's mean and computed lifetime sparseness,

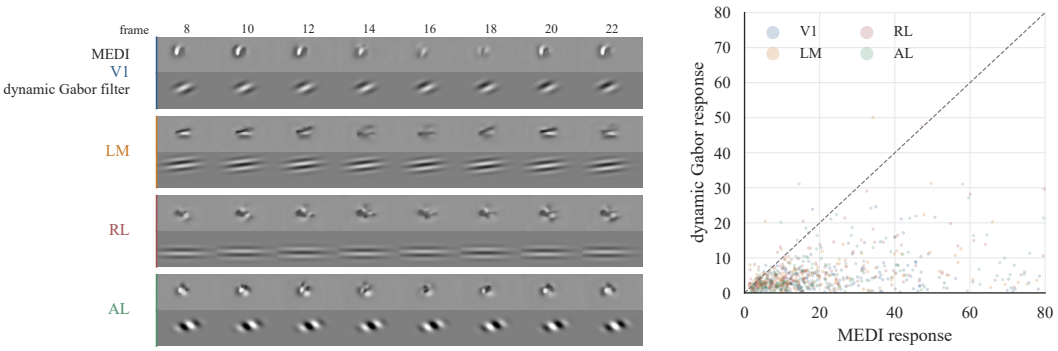
$$S_{i,a} = \frac{1 - (\bar{r}_{i,a})^2 / \overline{r^2}_{i,a}}{1 - 1/N_a}, \tag{1}$$

where a indexed the source-area MEDI pool, N_a was its number of MEDIs, $r_{i,a,k}$ was the normalized response of target readout i to MEDI k in pool a, $\bar{r}_{i,a} = N_a^{-1} \sum_k r_{i,a,k}$, and $\overline{r^2}_{i,a} = N_a^{-1} \sum_k r_{i,a,k}^2$. Lifetime sparseness measured response concentration across the stimulus set: values near zero indicated broadly distributed responses (Rolls & Tovee, 1995; Vinje & Gallant, 2000), whereas larger values indicated that a smaller subset of stimuli accounted for more of the response (Olshausen & Field, 2004). We computed $S_{i,a}$ separately for source pools V1, LM, RL, and AL and gave the four source-area scores equal weight, producing one area-balanced MEDI-set stimulus-selectivity score per target readout. Without this step, V1 would contribute many more stimuli simply because it has more readouts, and apparent concentration could reflect library composition rather than the readout's selectivity.

For transformation tolerance, each readout's own MEDI s_i was transformed and re-evaluated. For transformation T, we used response retention as the operational measure of tolerance:

$$\mathrm{ret}_{i,T} = \frac{r_i(T(s_i))}{r_i(s_i)}. \tag{2}$$

Retention was interpreted as a conditional model-output ratio, not a calibrated physiological fraction. Values near one indicated that the model retained the MEDI response under that specified change. We assessed contrast, rotation about the center of a MEDI-derived empirical ellipse, translation, complete temporal reversal, speed-up, and slow-down. Contrast used 0.25–2.0×, rotation used ±2 to ±90°, translation used 1–48 pixels in four directions on the 144×256 canvas (about 0.7–33% of the height and 0.4–18.8% of the width), with mean-value padding at exposed boundaries, speed-up used 1.25–2.0×, and slow-down used 0.5–0.8× speed. The complete parameterization, implementation details and representative transformed MEDI examples are in Appendix B.3. Interactive MEDI, transformation, and response examples are at `https://medi.gnefil.com/`.



(a) (b)

Figure 2: **MEDIs versus optimized dynamic Gabor stimuli.** (a) One example readout from each model area is shown as paired MEDI and dynamic-Gabor-filter keyframes. Frame indices are shown above the keyframes. Colored markers identify model V1, model LM, model RL, and model AL. (b) Paired responses are shown for all 800 readouts; the paired test uses the full cohort and is significant ($p < 10^{-100}$), while points are included only when both responses are ≤ 80 model-output units. Colors identify model areas and the dashed diagonal denotes equality. Response magnitudes are model-output units, not calibrated firing rates.

4 Results

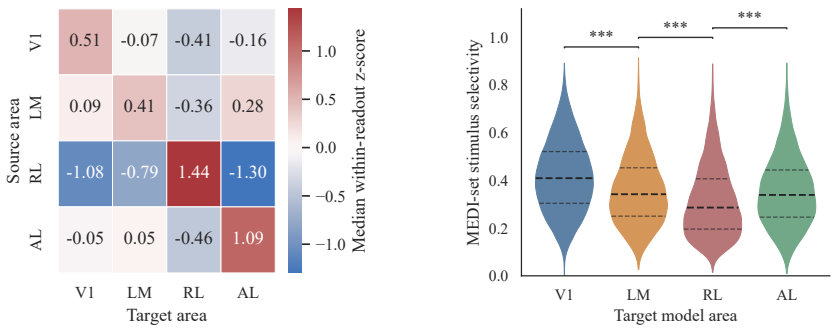
4.1 MEDIs elicited larger model responses than other stimulus baselines

We compared MEDIs with multiple stimulus baselines to assess whether neuron-specific dynamic optimization elicited stronger model responses than commonly used visual stimuli. Across the same frozen model, MEDIs produced the highest responses among all stimulus baselines tested, including static Gabor stimuli, optimized static stimuli (MESIs), drifting gratings, natural movies, and optimized dynamic Gabor stimuli; the complete set of baseline comparisons is reported in Appendix B.2. Because Gabor stimuli have been widely used as controlled reference stimuli in visual neuroscience, we focused the main-text comparison on dynamic Gabor stimuli, while reporting the other comparisons in the Appendix.

To make this comparison directly matched at the level of individual readouts, we optimized a dynamic Gabor stimulus separately for each readout using the procedure described in Methods. Figure 2a shows keyframes from one example readout in each model area, and Figure 2b summarizes the paired responses for the full 800-readout cohort. MEDI responses were significantly higher than dynamic-Gabor responses (paired one-sided Wilcoxon test, $p < 10^{-100}$). The statistical test uses all 800 readouts; for visualization, the plot displays only pairs for which both responses are ≤ 80 model-output units, preventing a small number of extreme predictions from compressing the main cloud. Additional MEDI, baseline, and response examples are available at `https://medi.gnefil.com/`.

4.2 Within-model-area response enrichment and model-area differences in MEDI-set stimulus selectivity

We next asked whether MEDIs optimized for one model area preferentially drove readouts from that same area, and how concentrated each readout's response was across the shared set. We grouped the MEDIs by the model area of the neuron whose response was optimized, giving four source-area MEDI groups; every target readout was evaluated on every group. For source–target enrichment, we averaged $\log(1+r)$ responses within each source-area group and z-scored the four group means within each target readout. A positive diagonal therefore means that a target readout responded more to MEDIs optimized for neurons in its own model area than to the other three groups; it is an enrichment summary, not a connectivity map. For MEDI-set stimulus selectivity, we normalized responses within each source-area group, computed lifetime sparseness as defined in Methods (Section 3.3), and



(a) (b)

Figure 3: **MEDI response structure.** (a) Rows are source-MEDI model areas and columns are target-readout model areas, displayed in the external-reference order V1–LM–RL–AL. Cell values are medians of within-target-readout z-scores of source-area mean-$\log(1 + r)$ responses. (b) Violin distributions show area-balanced MEDI-set stimulus selectivity across target readouts; short dashed lines mark the 25th, 50th, and 75th percentiles within each area. Brackets mark statistical comparisons between adjacent areas using two-sided Mann–Whitney U tests with Bonferroni correction across the three adjacent comparisons; "***" denotes corrected $p < 0.001$ and "n.s." denotes $p \geq 0.05$ (non-significant). Higher values indicate responses concentrated on fewer MEDIs. The response-matrix and selectivity summaries are model-predicted quantities, not measurements of synaptic connectivity.

averaged the four group scores with equal weight so that library size could not determine the result. Figure 3a shows source–target enrichment, whereas Figure 3b shows MEDI-set stimulus selectivity.

The matrix showed positive diagonal enrichment in all four model areas: V1 MEDIs preferentially drove V1 readouts, LM MEDIs preferentially drove LM readouts, and the AL and RL diagonals were especially strong, while cross-area blocks were generally weaker or negative. Median MEDI-set selectivity was the highest in model V1 (0.409), followed by model LM (0.342), model AL (0.339), and model RL (0.286), giving the empirical pattern V1 $>$ LM $\approx$ AL $>$ RL. Figure 3b retains the V1–LM–RL–AL display order corresponding to the established hierarchy, rather than sorting areas by their observed medians. Adjacent displayed-area two-sided Mann–Whitney comparisons with Bonferroni correction were significant for V1–LM, LM–RL, and RL–AL. A 5,000-resample readout bootstrap tested compatibility with either a monotonically increasing or decreasing profile along the partial order V1→LM/RL→AL, without assigning LM and RL different hierarchy levels. Neither direction was compatible with the observed profile in any bootstrap resample (Appendix B.4). Thus selectivity was area-organized but did not show a simple monotonic progression along the established hierarchy. The matrix therefore should not be interpreted as a connectivity map or as evidence of causal inter-area communication.

4.3 Transformation-specific tolerance across model areas

We then asked whether tolerance followed one model-area ordering across all transformations, or whether the ordering depended on the transformation being applied. We transformed each readout's own MEDI under six prespecified conditions—contrast, rotation, translation, temporal reversal, speed-up, and slow-down—and summarized response retention across the tested parameters within each condition as described in Methods and Appendix B.3. Figure 4 reports the four model-area median profiles with standard-deviation bands. Its adjacent displayed-order comparisons use the same readout-level summaries as the plotted medians: all three contrasts were significant for contrast, rotation, temporal reversal, and speed-up; translation and slow-down showed significant V1–LM and LM–RL differences but not RL–AL. The empirical profiles increased across the displayed V1–LM–RL–AL order for contrast, rotation, and speed-up; increased through RL for translation; followed V1 $>$ LM and LM $<$ RL $<$ AL for temporal reversal; and decreased through RL for slow-down. Thus the profiles differed across transformations, including relatively greater retention in model RL under strong spatial transformations, relatively greater retention in model V1 under slow-down, and relatively less retention in model V1 under strong speed-up.

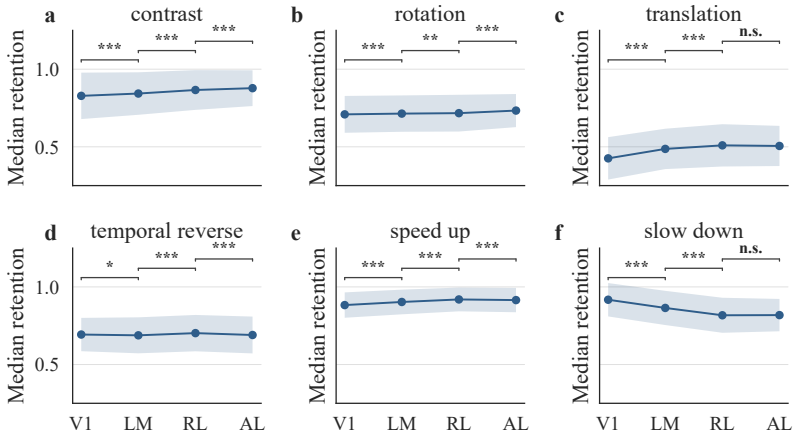


Figure 4: **Transformation-specific tolerance across 104,171 readouts.** Connected blue points are area medians in the external-reference display order V1–LM–RL–AL (LM and RL are the shared intermediate level), and translucent bands show one standard deviation across readouts. Each readout is first summarized by the median across repeated entries at each tested parameter and, for speed conditions, across speed factors within that condition. Adjacent displayed-order comparisons were assessed using bootstrap percentile intervals and two-sided readout-level Mann–Whitney tests, with Holm correction across all 18 transformation-by-pair tests; "***", "**", "*", and "n.s." denote corrected $p < 0.001$, $p < 0.01$, $p < 0.05$, and $p \geq 0.05$, respectively.

We next tested whether each profile was compatible with a simple monotonic progression along the established partial-order reference V1→LM/RL→AL, treating LM and RL as one intermediate level. A readout bootstrap evaluated increasing compatibility when model V1 was no greater than both intermediate areas and both intermediate areas were no greater than model AL; decreasing compatibility used the reverse inequalities. The proportion of bootstrap resamples satisfying the increasing constraints was 1.00 for contrast and rotation, 0.119 for translation, and 0.013 for speed-up, and zero for temporal reversal and slow-down. The corresponding proportion for a decreasing profile was 0.292 for slow-down and zero for the other transformations. No single monotonic organization therefore consistently described tolerance across all six transformations: some profiles were compatible with the established reference, whereas others were transformation-specific or incompatible with either direction. The parameter-resolved profiles and hierarchy-consistency output are provided in Appendix C.1. We additionally assessed robustness to readout-quality selection using the V1-anchored analysis described in Appendix C.2; this served only as a robustness check rather than a test of global area ordering.

4.4 Selectivity–tolerance coupling depended on the transformation

We next asked whether readouts that were selective for a narrower portion of the shared MEDI set also showed systematically different tolerance. This question tests whether the selectivity–tolerance relationship is transformation-specific or whether one universal association describes all transformations and strengths. We computed Spearman correlations between area-balanced MEDI-set stimulus selectivity and retention across readouts within each model area. For the main figure, we show one prespecified stress per transformation: the parameter with the lowest median retention after pooling all readouts—contrast $0.25\times$, rotation $90°$, translation 48 pixels, temporal reverse, speed-up $2\times$, and slow-down $0.5\times$. These parameters focus the comparison on regimes with appreciable retention variation; at weaker parameters, retention can be compressed near one.

At the selected stresses, readouts whose responses were concentrated on fewer MEDIs tended to lose more response under the corresponding transformation. The associations were nearly unchanged when each target readout's own MEDI was excluded from its selectivity calculation (Appendix D.1). Figure 5 shows the selected-stress associations. Keeping the transformations separate yields six testable model predictions that can be assessed with matched in vivo presentations.

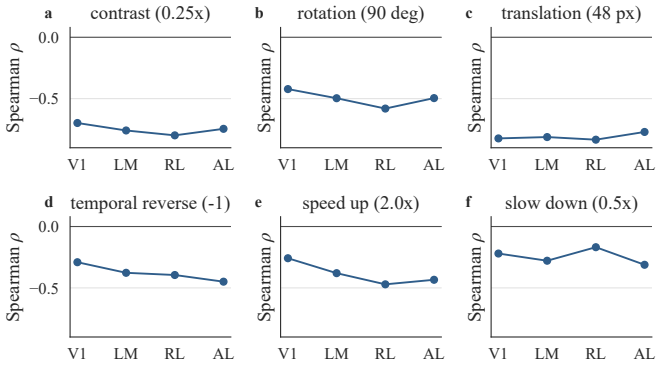


Figure 5: **Selectivity–tolerance association at the lowest median retention.** Each subplot title gives the transformation and selected stress parameter. Connected blue points show Spearman ρ for V1, LM, RL, and AL; the gray line marks zero. Selectivity is area-balanced lifetime sparseness and tolerance is response retention.

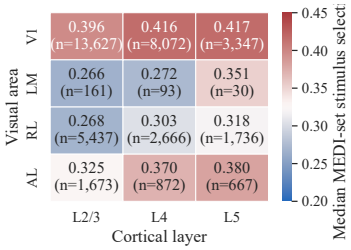


Figure 6: **MEDI-set stimulus selectivity by cortical layer.** Rows are visual areas and columns are cortical layers. Cell color and the first number encode median selectivity; parentheses give labelled-readout counts. LM has fewer labelled readouts, especially in L5 ($n = 30$).

The complete parameter-by-parameter correlations are given in Appendix D.2. A deterministic nested split separated stress selection from estimation; Appendix D.3 reports held-out estimates, bootstrap intervals, and correction across the 48 split-by-transformation-by-area tests. Both splits recovered negative associations at the selected stresses, with slow-down weakest. Together, these analyses support a transformation-specific selectivity–tolerance relationship: the association remained negative across the six tested transformations, but its magnitude varied with transformation and parameter rather than defining a universal law or an additional area hierarchy.

4.5 MEDI-set stimulus selectivity was lower in L2/3 within every model area

We then asked whether the area-level selectivity pattern was accompanied by systematic differences across cortical layers within each model area. Layer labels inherited from the corresponding recorded neurons were available for 38,381 readouts. Within this labelled sample, median MEDI-set stimulus selectivity was lower in L2/3 than in L4 and L5 in all four model areas (Figure 6). A Kruskal–Wallis test supported a layer difference within every area after Benjamini–Hochberg correction ($q < 0.01$ for each area). Pairwise tests showed lower L2/3 selectivity than L5 in every area and lower L2/3 selectivity than L4 in V1, RL, and AL (all corrected $q < 0.01$); the smaller LM L2/3–L4 comparison was not significant ($q = 0.739$).

These tests support the repeated direction of the layer medians without implying a universal laminar mechanism. LM contains only 161 labelled readouts in L2/3, 93 in L4, and 30 in L5, so its estimates are less precise than those for V1, RL, and AL. Appendix E.1 separately reports response retention by area, cortical layer, and transformation; those crossings do not reduce to one layer ordering.

4.6 High temporal-reverse retention occurs for both near-static and strongly changing MEDIs

Temporal reversal produced distinct retention profiles across readouts, including cases with very high retention despite substantial stimulus changes over time. We therefore asked whether high temporal-reverse retention primarily reflected nearly static MEDIs, or whether some readouts preserved their response despite substantial forward–reverse change. We quantified stimulus change as the mean pixelwise difference between each original MEDI frame and its temporally reversed counterpart. Among the top decile of retention, values below the within-area third quartile defined near-static candidates, whereas values above the top quartile defined a large-change category. MEDIs in the top retention decile were more static than the remaining MEDIs in every area, with median distances of 0.033 versus 0.044 in V1, 0.030 versus 0.045 in LM, 0.026 versus 0.042 in RL, and 0.028 versus 0.045 in AL. Nevertheless, the top-retention group contained both many near-static readouts (V1 6,297; LM 1,203; RL 1,061; AL 388) and a smaller set combining high retention with large pixel

change (V1 1,253; LM 126; RL 63; AL 28). Appendix E.2 shows the full screen and representative MEDIs.

These observations support two descriptive interpretations of high reversal retention: it can arise when temporal reversal produces little stimulus change, or when substantial stimulus change is accompanied by preserved response. The latter remains phenomenological, because mean response and pixel distance alone cannot distinguish order-invariant temporal integration from shared framewise energy, saturation, or other confounds. One testable hypothesis is that the model may depend more on framewise texture or shape than on temporal order. Replication with independently optimized MEDIs would test whether the large-change category generalizes beyond this fitted system.

5 Conclusion and Discussion

Conclusion. The principal findings of this model-based analysis include: (i) MEDIs elicited larger model responses than dynamic and static Gabor stimuli, MESIs, drifting gratings, and natural movies; (ii) the shared MEDI set showed positive within-model-area enrichment, and MEDI-set selectivity was the highest in model V1 and the lowest in model RL, with model LM and model AL intermediate and close to one another rather than forming a single monotonic sequence; (iii) tolerance was transformation-specific: model V1 retained the least response under contrast, translation, and speed-up and the most under slow-down, whereas model RL retained relatively more response under strong spatial transformations, with different transformations producing different area profiles; (iv) selectivity and tolerance were negatively associated at the selected strength of every transformation, but the association varied with transformation and parameter; and (v) layer-labelled readouts showed lower MEDI-set selectivity in L2/3 than in L4 and L5 within each model area.

Findings consistent with animal experiments. Several conclusions show qualitative correspondence with prior animal experiments. Conclusion (i) is consistent with in vivo evidence that optimized stimuli can drive stronger neuronal responses than conventional Gabor-like stimuli (Walker et al., 2019). Conclusion (iv) is consistent with reported negative selectivity–tolerance associations in macaque IT and V4 (Zoccolan et al., 2007; Sharpee et al., 2013). The dependence of this association on transformation and parameter is also consistent with experiments showing that tolerance varies with image statistics preserved by a transformation (Ziemba et al., 2016; Freeman & Simoncelli, 2011). These parallels provide qualitative correspondence rather than validation of the model-derived area patterns.

Findings that differ from prior reports. Conclusion (iv) differs from cross-stage analyses reporting joint increases in selectivity and tolerance (Rust & DiCarlo, 2010), and from within-area analyses reporting no direct association between them (Rust & DiCarlo, 2012). Those studies compared cortical stages using object or natural-image stimulus sets, whereas conclusion (iv) reflects within-area variation across model readouts using optimized dynamic stimuli. The difference might be attributed to the different stimuli used in experiments.

Findings that generate testable predictions. Our protocol yields four testable in vivo predictions: (i) MEDIs should outperform dynamic and static Gabor stimuli, MESIs, drifting gratings, and natural movies; (ii) area-matched MEDIs should enrich responses and selectivity should differ across areas; (iii) neurons should show the predicted transformation-specific tolerance profile, including the contrast/translation/speed-up versus slow-down dissociation; and (iv) selectivity and tolerance should show the predicted negative associations and layer pattern. Matched stimulus presentations and chronic imaging can test these predictions in the same neurons.

Limitations and future experiments. Three limitations define the scope. First, MEDIs are optimized and evaluated with the same fitted model and may exploit model-specific errors; checkpoint (Appendix A.2) and latent-seed (Appendix A.3) controls are robustness checks, not independent validation. Second, generator prior, regularization, behavioral covariates, and incomplete layer annotations constrain accessible stimulus–response geometry. Third, readouts are associated with recordings from one mouse, so stability is readout-level within this fitted system, not population-level. Independent response fitting, matched in vivo MEDI presentations, and replication across animals would strengthen biological claims.

ACKNOWLEDGMENTS

We thank the MICrONS consortium and the authors of the neural-activity foundation model for making the data, model, and analysis resources publicly available.

Reproducibility statement. Code and released analysis materials are available at the anonymous repository `https://anonymous.4open.science/r/mouse_medi_reproducibility-081C/`. The appendix is organized into five parts: robustness and sensitivity analyses (Appendices A.1, A.2, and A.3); Supplementary Methods (Appendices B.1–B.8); quality and areal tolerance (Appendices C.1 and C.2); and selectivity–tolerance analyses (Appendices D.1, D.2, and D.3); and supplementary validation (Appendices E.1 and E.2). The repository includes the released readout manifest, compact seed-summary inputs, derived tables, figure-generation scripts, and an executable regeneration entry point. The dense response matrix, licensed foundation-model weights, and 104,171 optimized videos are intentionally regenerated from the documented external-artifact contract rather than bundled. All reported bootstrap intervals quantify stability over correlated model readouts, not independent animals.

AI use statement. Generative AI tools were used to assist with language editing and code organization. The authors reviewed all generated text and code, verified numerical claims against the analysis outputs, and take responsibility for the final manuscript.

REFERENCES

Mark L. Andermann, Aaron M. Kerlin, Demetris K. Roumis, et al. Functional specialization of mouse higher visual cortical areas. *Neuron*, 72:1025–1039, 2011. doi: 10.1016/j.neuron.2011.11.013.

Mohammad Bashiri, Luca Baroni, Ján Antolík, and Fabian H. Sinz. Learning and aligning single-neuron invariance manifolds in visual cortex. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=kbjJ9ZOakb.

Ollin Boer Bohan. Tiny autoencoder for stable diffusion (taesd). https://github.com/madebyollin/taesd, 2023. Commit dc25abb; accessed 2026-01-29.

Max F. Burg, Thomas Zenkel, Michaela Vystrčilová, Jonathan Oesterle, Larissa Höfling, Konstantin F. Willeke, Jan Lause, Sarah Müller, Paul G. Fahey, Zhiwei Ding, Kelli Restivo, Shashwat Sridhar, Tim Gollisch, Philipp Berens, Andreas S. Tolias, Thomas Euler, Matthias Bethge, and Alexander S. Ecker. Most discriminative stimuli for functional cell type clustering. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=9W6KaAcYlr.

Zhiwei Ding, Dat T. Tran, Kayla Ponder, et al. Functional bipartite invariance in mouse primary visual cortex receptive fields. *Nature Neuroscience*, 29:851–863, 2026. doi: 10.1038/s41593-026-02213-3.

Rinaldo D. D'Souza, Quanxin Wang, et al. Hierarchical and nonhierarchical features of the mouse visual cortical network. *Nature Communications*, 13:503, 2022. doi: 10.1038/s41467-022-28035-y.

Thomas Fel, Thibaut Boissin, Victor Boutin, Agustin Picard, Paul Novello, Julien Colin, Drew Linsley, Tom Rousseau, Rémi Cadène, Lore Goetschalckx, Laurent Gardes, and Thomas Serre. Unlocking feature visualization for deeper networks with magnitude constrained optimization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 37813–37826, 2023.

Jeremy Freeman and Eero P. Simoncelli. Metamers of the ventral stream. *Nature Neuroscience*, 14:1195–1201, 2011. doi: 10.1038/nn.2889.

Emmanouil Froudarakis, Uri Cohen, Maria Diamantaki, et al. Object manifold geometry across the mouse cortical visual hierarchy. *eLife*, 2026. doi: 10.7554/eLife.110287.

Xu Han, Ben Vermaercke, and Vincent Bonin. Diversity of spatiotemporal coding reveals specialized visual processing streams in the mouse cortex. *Nature Communications*, 13:3249, 2022. doi: 10.1038/s41467-022-29656-z.

Julie A. Harris, Stefan Mihalas, et al. Hierarchical organization of cortical and thalamic connectivity. *Nature*, 575:195–202, 2019. doi: 10.1038/s41586-019-1716-z.

Jay Hegdé and David C. Van Essen. Stimulus dependence of disparity coding in primate visual area v4. *Journal of Neurophysiology*, 93(1):620–626, 2005. doi: 10.1152/jn.00039.2004.

Lukas Höfling et al. A chromatic feature detector in the retina signals visual context changes. *eLife*, 13:86860, 2024. doi: 10.7554/eLife.86860.

James H. Marshel, Marina E. Garrett, Ian Nauhaus, and Edward M. Callaway. Functional specialization of seven mouse visual cortical areas. *Neuron*, 72:1040–1054, 2011. doi: 10.1016/j.neuron.2011.12.004.

MICrONS Consortium. Functional connectomics spanning multiple areas of mouse visual cortex. *Nature*, 640:435–447, 2025. doi: 10.1038/s41586-025-08790-w.

Kevin A. Murgas, Ashley M. Wilson, Valerie Michael, and Lindsey L. Glickfeld. Unique spatial integration in mouse primary visual cortex and higher visual areas. *Journal of Neuroscience*, 40(9):1862–1873, 2020. doi: 10.1523/JNEUROSCI.1997-19.2020.

Anh Nguyen, Jason Yosinski, Yoshua Bengio, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016a.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. In *Workshop on Visualization for Deep Learning, ICML*, 2016b.

G. Okazawa, S. Tajima, and H. Komatsu. Neural representations of shape in macaque v4. *Nature Communications*, 6:6216, 2015. doi: 10.1038/ncomms7216.

Bruno A. Olshausen and David J. Field. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4):481–487, 2004. doi: 10.1016/j.conb.2004.07.007.

Eugenio Piasini, Liviu Soltuzu, Paolo Muratore, Riccardo Caramellino, Kasper Vinken, and Hans Op de Beeck. Temporal stability of stimulus representation increases along rodent visual cortical hierarchies. *Nature Communications*, 12:4448, 2021. doi: 10.1038/s41467-021-24456-3.

Pawel A. Pierzchlewicz, Konstantin F. Willeke, Arne Nix, Pavithra Elumalai, Kelli Restivo, Tori Shinn, Cate Nealley, Gabrielle Rodriguez, Saumil Patel, Katrin Franke, Andreas S. Tolias, and Fabian H. Sinz. Energy guided diffusion for generating neurally exciting images. In *Advances in Neural Information Processing Systems*, volume 36, pp. 32574–32601, 2023. doi: 10.52202/075280-1414.

Carlos R. Ponce, Will Xiao, Peter F. Schade, et al. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, 177:999–1009, 2019. doi: 10.1016/j.cell.2019.04.005.

Edmund T. Rolls and Martin J. Tovee. Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *Journal of Neurophysiology*, 73(2):713–726, 1995. doi: 10.1152/jn.1995.73.2.713.

Nicole C. Rust and James J. DiCarlo. Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area v4 to it. *Journal of Neuroscience*, 30(39):12978–12995, 2010. doi: 10.1523/JNEUROSCI.0179-10.2010.

Nicole C. Rust and James J. DiCarlo. Balanced increases in selectivity and tolerance produce constant sparseness along the ventral visual stream. *Journal of Neuroscience*, 32(30):10170–10182, 2012. doi: 10.1523/JNEUROSCI.6125-11.2012.

Tatyana O. Sharpee, Minjoon Kouh, and John H. Reynolds. Trade-off between curvature tuning and position invariance in visual area v4. *PNAS*, 110(28):11618–11623, 2013. doi: 10.1073/pnas.1217479110.

Joshua H. Siegle, Xiaoxuan Jia, et al. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, 592:86–92, 2021. doi: 10.1038/s41586-020-03171-x.

Yingtian Tang, Sogand Salehi, Ming Zhou, Amir Zamir, Leyla Isik, and Martin Schrimpf. Nevo: Neural-guided evolutionary video synthesis for dynamic visual selectivity, 2026.

Ben Vermaercke, Florian J. Gerich, Ellen Ytebrouck, Lutgarde Arckens, Hans P. Op de Beeck, and Gert Van den Bergh. Functional specialization in rat occipital and temporal visual cortex. *Journal of Neurophysiology*, 112(8):1963–1983, 2014. doi: 10.1152/jn.00737.2013.

William E. Vinje and Jack L. Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276, 2000. doi: 10.1126/science.287.5456.1273.

Edgar Y. Walker, Fabian H. Sinz, Erick Cobos, et al. Inception loops discover what excites neurons most using deep predictive models. *Nature Neuroscience*, 22:2060–2065, 2019. doi: 10.1038/s41593-019-0517-x.

Binxu Wang and Carlos R. Ponce. Tuning landscapes of the ventral stream. *Cell Reports*, 40(8):111595, 2022. doi: 10.1016/j.celrep.2022.111595.

Binxu Wang and Carlos R. Ponce. On the level sets and invariance of neural tuning landscapes. In *Proceedings of the 1st NeurIPS Workshop on Symmetry and Geometry in Neural Representations*, volume 197, pp. 278–300, 2023.

Eric Y. Wang, Paul G. Fahey, Zhuokun Ding, et al. Foundation model of neural activity predicts response to new stimulus types. *Nature*, 640:470–477, 2025. doi: 10.1038/s41586-025-08829-y.

Will Xiao and Gabriel Kreiman. Xdream: Finding preferred stimuli for visual neurons using generative networks and gradient-free optimization. *PLoS Computational Biology*, 16:e1007973, 2020. doi: 10.1371/journal.pcbi.1007973.

Chun-I Yeh, Dajun Xing, Patrick E Williams, and Robert M Shapley. Stimulus ensemble and cortical layer determine v1 spatial receptive fields. *Proceedings of the National Academy of Sciences*, 106 (34):14652–14657, 2009. doi: 10.1073/pnas.0907406106.

Yiyi Yu, Jeffrey N. Stirman, Christopher R. Dorsett, and Spencer L. Smith. Selective representations of texture and motion in mouse higher visual areas. *Current Biology*, 32:2810–2820.e5, 2022. doi: 10.1016/j.cub.2022.04.091.

Kalathupiriyan A. Zhivago and S. P. Arun. Selective it neurons are selective along many dimensions. *Journal of Neurophysiology*, 115(3):1512–1520, 2016. doi: 10.1152/jn.01151.2015.

Corey M. Ziemba, Jeremy Freeman, J. Anthony Movshon, and Eero P. Simoncelli. Selectivity and tolerance for visual texture in macaque v2. *PNAS*, 113:E3140–E3149, 2016. doi: 10.1073/pnas.1510847113.

Davide Zoccolan, Minjoon Kouh, Tomaso Poggio, and James J. DiCarlo. Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *Journal of Neuroscience*, 27(45): 12292–12307, 2007. doi: 10.1523/JNEUROSCI.1897-07.2007.

A SUPPLEMENTARY ROBUSTNESS AND SENSITIVITY ANALYSES

A.1 ROBUSTNESS TO AUTOENCODER CHOICE

To test whether latent-space optimization depends on the autoencoder, we compared five models spanning distinct training domains and representational properties: TAESD (Tiny Autoencoder for Stable Diffusion), a highly compressed natural-image autoencoder (Bohan, 2023); Zeroscope VAE, trained on video frames; SD-VAE, the canonical Stable Diffusion image autoencoder; SDXL-VAE, its higher-capacity successor; and a Medical VAE trained on brain-MRI data. Each model was used with the same target readouts, response objective, and optimization schedule after matching its decoded output to the foundation-model input size. Despite differences in training data and reconstruction fidelity, all five produced qualitatively similar MEDI structure and the same localized high-response patterns. TAESD was used for the main experiments because its compact decoder makes iterative optimization substantially faster.

A.2 ALTERNATIVE-CHECKPOINT SENSITIVITY

We re-evaluated the original MEDIs and transformed variants for 1,000 stratified readouts using the released orientation-finetuned checkpoint (`model_ori_fine_tuned`). This is not an independent model: it shares the foundation-model core, so it is a checkpoint/readout sensitivity analysis rather than an external validation. Nevertheless, original MEDI responses agree with Spearman $\rho = 0.944$, and readout-by-family retention summaries agree at $\rho = 0.911$–0.958 across contrast, rotation, translation, temporal reverse, and the checkpoint analysis's temporal-timing family. Because that archived analysis reports temporal timing as one family, it is not used to infer a joint speed-up/slow-down effect; the primary analysis keeps those two directions separate. The alternative checkpoint shifts absolute median retention, so this control supports stability of rank structure but not calibration of the effect size or validity in vivo. The reported rank summaries are retained in the manuscript analysis record.

A.3 INDEPENDENT LATENT-SEED ROBUSTNESS

We regenerated and re-evaluated a predeclared subset of 64 readouts balanced across visual areas and cortical-layer groups (16 each from V1, LM, RL, and AL; four each from L2/3, L4, L5, and the unassigned group) with production seed 42 and independent latent seed 1042. Original responses agreed at Spearman $\rho = 0.997$; condition-level retention correlations ranged from $\rho = 0.882$ to 0.998, and the 24 area-by-condition median profiles agreed at $\rho = 0.997$. Median frame-aligned pixel correlation was 0.966 (RMSE 0.0109), confirming that the generated videos differ while their extracted response and retention metrics remain highly similar. This small, balanced cohort probes optimization stability only; it is not used to estimate or validate a brain-area gradient. Summary inputs are supplied as `figures/inputs/cross_seed_*.csv` together with Figure 7. The same evaluator and model core are used.

B SUPPLEMENTARY METHODS

B.1 MEDI SYNTHESIS OBJECTIVE AND IMPLEMENTATION

For target readout i, let $z_{1:T}$ be a sequence of latent tensors and $x_t = D(z_t)$ the decoded frames. The decoder is TAESD with four latent channels and a 36×64 latent grid; decoded frames are bilinearly resized to 144×256, converted to luminance with ITU-R BT.601 weights $(0.299, 0.587, 0.114)$, clipped to $[-1, 1]$, and mapped to $[0, 1]$ before entering the frozen model. We minimize

$$\mathcal{L} = -\lambda_{\text{act}} \log(\epsilon + T^{-1} \sum_t r_i(x_t)) + \lambda_{\text{sp}} \mathcal{L}_{\text{sp}} + \lambda_{\text{pix}} \mathcal{L}_{\text{pix}} + \lambda_{\text{lat}} \mathcal{L}_{\text{lat}}, \tag{3}$$

where $\epsilon = 10^{-4}$, $\mathcal{L}_{\text{sp}}$ is the mean squared horizontal and vertical latent-neighbor difference, $\mathcal{L}_{\text{pix}}$ is the mean squared adjacent-frame difference after decoding, and $\mathcal{L}_{\text{lat}}$ is the mean squared adjacent-frame latent difference. The production settings are $T = 60$, latent shape $(1, 4, 1, 36, 64)$, $\lambda_{\text{act}} = 3$, $\lambda_{\text{sp}} = 500$, $\lambda_{\text{pix}} = 5000$, and $\lambda_{\text{lat}} = 2000$. The latent is initialized as $0.5 + 0.02\,\mathcal{N}(0, 1)$, optimized by SGD with learning rate 10, momentum 0.9, gradient-norm clipping at 0.1, cosine decay to 1,

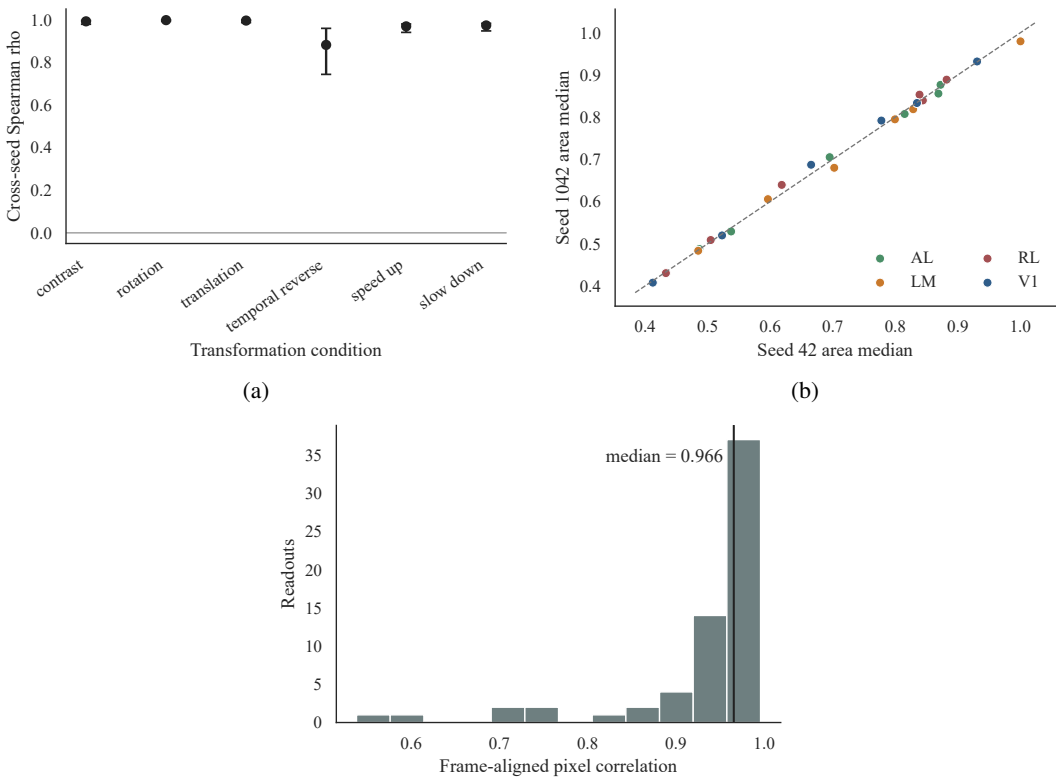


Figure 7: **Independent latent-seed replication on a stratified subset.** Seed 42 and seed 1042 regenerate the same 64-readout cohort under identical production settings. (a) Black points show condition-level cross-seed Spearman correlations across readouts; vertical bars are 95% readout-bootstrap intervals, and the gray horizontal line marks zero. (b) Each colored point is one of the 24 combinations of visual area and transformation condition, with its coordinates giving the median retention for seed 42 (horizontal axis) and seed 1042 (vertical axis); the dashed gray identity line denotes equal medians across seeds. (c) Histogram of the 64 frame-aligned pixel correlations between paired MEDIs; the solid vertical line and text annotation mark the median. The extracted response and retention metrics are highly similar despite non-identical stimuli. This balanced small-cohort analysis tests optimization-basin sensitivity only; it does not estimate a brain-area gradient or provide independence from the shared model core.

and in-place clipping to $[-5, 5]$. At every update, independent Gaussian latent noise with standard deviation 0.05 is added before decoding; the target objective is the mean of the model outputs over all 60 frames. The model is reset before each forward pass, and behavioral/position covariates are zero tensors, matching the released generation code.

The sequence is optimized with 15-frame windows and 8-frame overlap. The first window is initialized near a gray background using the released latent initialization. For each subsequent window, the overlap is copied from the preceding optimized window and the new frames start from the same gray-background initialization; independent Gaussian latent noise is injected before every update. Shared frames are retained once when windows are concatenated into the final 60-frame sequence. Random translation, scaling, and blur are disabled in production. A pixel-space comparison is shown only to document why the latent prior is used: direct frame optimization produced high-frequency artifacts and unstable temporal structure under comparable smoothness settings. Figure 8 shows the matched four-readout visual comparison.

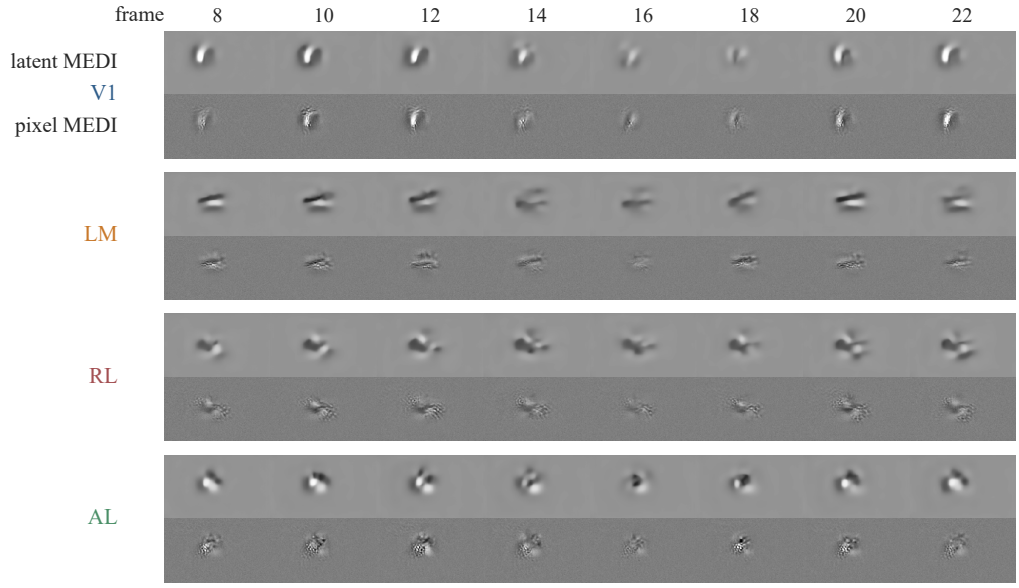


Figure 8: **Latent- versus pixel-space MEDI optimization.** For one readout from each visual area, the upper row is the released latent-space MEDI and the lower row is a direct pixel-space optimization evaluated by the same frozen model. Columns show matched temporal samples from each 60-frame video. The pixel-space examples exhibit high-frequency texture and less coherent temporal structure, motivating the latent prior used for production MEDIs.

B.2 Baseline stimulus generation and evaluation

All baselines use the identical frozen foundation-model checkpoint, preprocessing, session/scan/readout mapping, and response extraction used for MEDI. Inputs are converted to the model's single-channel 144×256 format. For every 60-frame stimulus we discard the first 10 warm-up frames and report the mean of the remaining 50 model outputs. Thus values are paired within readout (exact `session/scan_idx/readout_id` keys) and are in model-output units rather than calibrated firing rates. The released comparison cohort contains 200 randomly selected readouts per area (V1, LM, RL, AL), with the random selection recorded in the accompanying CSV files.

Dynamic Gabor stimulus. Dynamic Gabor stimuli are parameterized jointly by spatial orientation $\theta \in [-2\pi, 2\pi]$, Gaussian width $\sigma \in [0.13, 0.25]$, wavelength $\Lambda \in [0.2, 2.0]$, phase $\psi \in [-\pi, \pi]$, aspect ratio γ, contrast, normalized center coordinates in $[-0.55, 0.55]^2$, temporal frequency $f_t \in [-0.32, 0.32]$, and image velocity $v \in [-0.006, 0.006]$ (normalized image units per frame); aspect ratio and contrast use the validated implementation bounds. Parameters are optimized jointly for all 60 frames against the target readout, with the released learning rate, Adam optimizer, iteration count, and parameter projection; no temporal resampling or auxiliary transformation is introduced. The final 60-frame video is regenerated from the optimized parameters and evaluated with the common 10-frame warm-up rule above. This is the only baseline with time-varying Gabor parameters.

Static Gabor stimulus. Static Gabor stimuli use exactly the same spatial equations, bounds, initialization, optimizer, objective, and 60-frame evaluation as dynamic Gabor stimuli, but set $f_t = v = 0$ and replicate one spatial frame throughout the clip. Consequently the comparison changes only temporal degrees of freedom and not stimulus size, contrast, or evaluator.

MESI. A MESI is a single optimized image generated with the static latent pipeline (four latent channels decoded to 144×256 pixels). We use the released 20-frame, 10-iteration optimization configuration, learning rate 10, activation objective, spatial regularizer, Gaussian latent noise, and the same latent initialization and model preprocessing as the synthesis code. The resulting image is

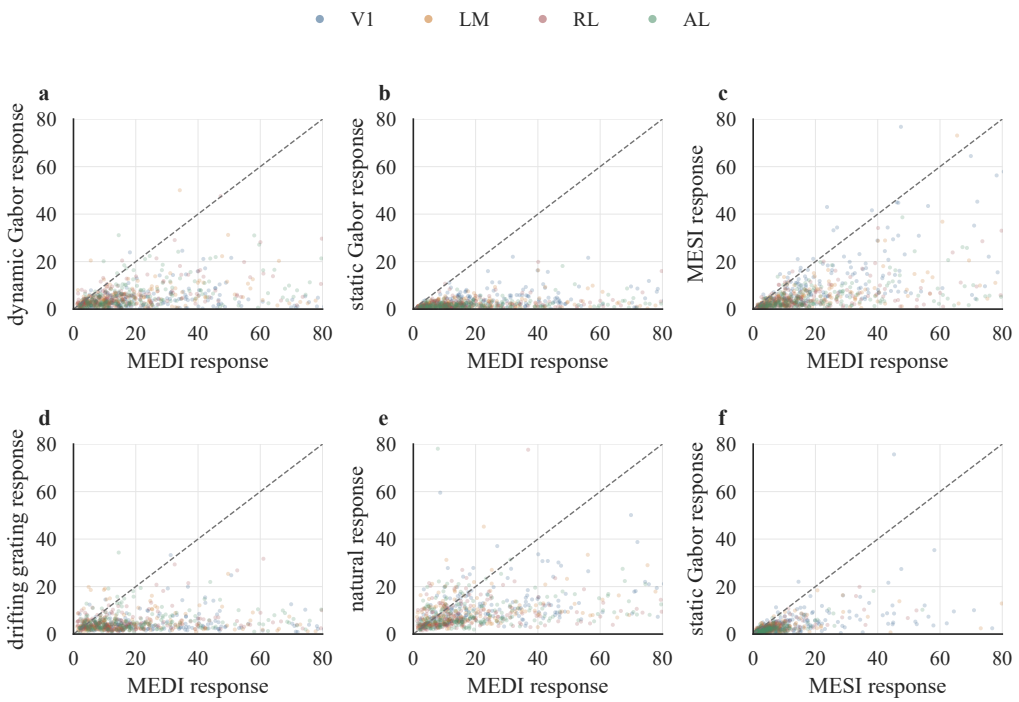


Figure 9: **MEDI and optimized-stimulus baseline comparisons.** The first five panels are paired scatters with MEDI response on the horizontal axis and one baseline on the vertical axis; the final panel directly compares MESI on the horizontal axis with static Gabor on the vertical axis. For display, both responses are limited to ≤ 80 model-output units. Colors identify V1, LM, RL, and AL, and dashed diagonals denote equality. All six paired comparisons are significant at $p < 3 \times 10^{-77}$. The drifting-grating response is the maximum over its complete tested grid. Response magnitudes are model-output units, not calibrated firing rates.

repeated for 60 evaluation frames, after which the first 10 frames are discarded and the remaining 50 are averaged. Repetition is an evaluation convention only; MESI has no temporal degrees of freedom.

Drifting grating. We probe drifting sinusoidal gratings on a factorial grid of 16 directions equally spaced over 360° (22.5° increments), five exponentially spaced temporal frequencies (0.5, 1, 2, 4, and 8 Hz), and six exponentially spaced spatial frequencies (0.01, 0.02, 0.04, 0.08, 0.16, and 0.32 cycles per pixel; approximately 0.02–0.64 cycles per degree in the foundation-model setup). This gives $16 \times 5 \times 6 = 480$ direction–TF–SF combinations per readout at unit contrast. Every condition is rendered as a 60-frame clip and evaluated by the same frozen model; after discarding the first 10 frames, the response is averaged over the remaining 50 frames. For each readout we retain the combination with the highest response. The released table records that maximum together with the preferred direction, temporal frequency, and spatial frequency, while the 480-condition grid and its parameterization are preserved in the generation metadata.

Natural movie. Natural controls use the foundation-model training stimulus and response arrays (`stimulus.npy` and `responses.npy`); these external arrays are not bundled in the repository and their location is specified in the data README. For each readout and each independent 300-frame clip, we scan contiguous 60-frame windows in `responses.npy`, select the highest-response window, then re-run the corresponding frames from `stimulus.npy` through the frozen model and average the final 50 outputs after the 10-frame warm-up. Windows are never concatenated across clips. This provides a data-driven movie control without optimizing pixels or latents.

The five controls are compared with MEDI only within readout. Figure 9 shows all available paired rows. MEDI responses exceed each baseline, and MESI responses exceed static-Gabor responses (paired one-sided Wilcoxon tests, all $p < 3 \times 10^{-77}$). The direct MESI–static-Gabor comparison

reproduces the principal optimized-stimulus comparison reported by Walker et al. for mouse V1 (Walker et al., 2019). These comparisons remain conditional on the shared model and should not be read as claims about calibrated firing rates.

Additional MEDI, baseline, transformation, and response examples are available at `https://medi.gnefil.com/`.

B.3 TRANSFORMATION DEFINITIONS

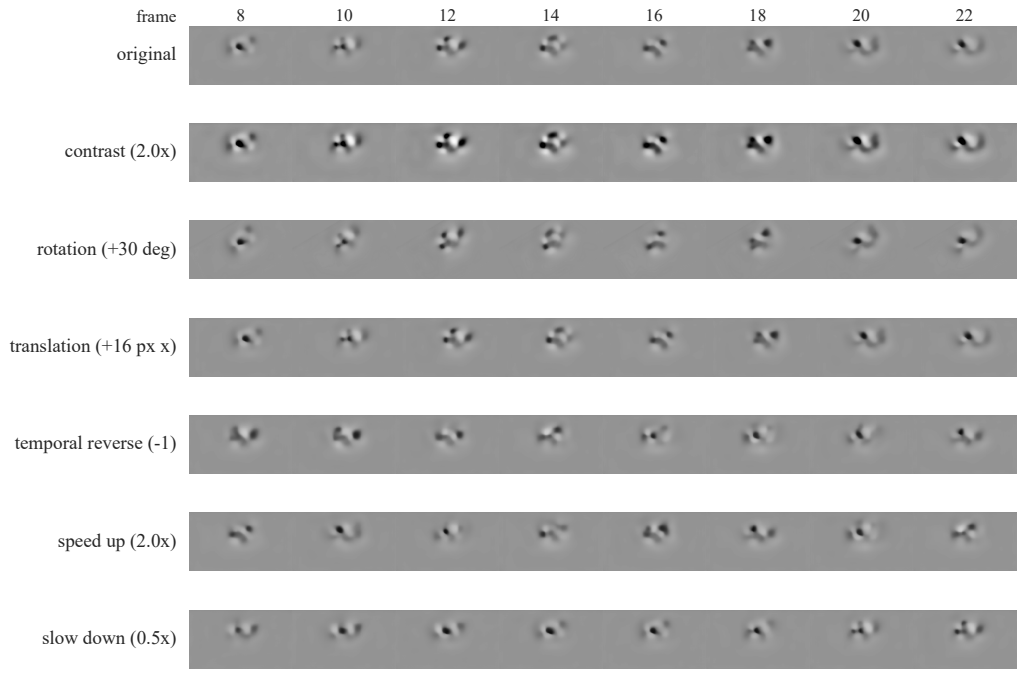


Figure 10: **MEDI and transformed examples.** Rows show one original MEDI and the six transformations used in the analysis. The "frame" header identifies the centered keyframe indices shown above the columns. These images define the visual interventions and are not additional statistical samples. Additional transformation examples are available at `https://medi.gnefil.com/`.

Figure 10 shows representative transformed MEDIs. For each target's own MEDI, we generate the following transformed sequences. The released parameter table retains one retention value for every readout and every listed parameter; the response and original-response fields used to form that ratio belong to the transformation-evaluation export, whereas the compact released table stores only the resulting retention. Condition-level medians are computed only after this per-readout parameter table is formed. Rotations use bicubic interpolation; translations are integer shifts with mean-value boundary filling. All spatial outputs are clipped to the model input range. speed-up and slow-down use linear temporal resampling of the complete clip with variable output duration, retaining both endpoints and never wrapping from the last frame to the first. They therefore implement non-cyclic playback-rate interventions. The variable-duration speed results are summarized in the accompanying analysis data. Temporal reversal is analyzed separately and never pooled with them.

- **Contrast:** each frame's deviation from its spatial mean is multiplied by $0.25, 0.5, 0.75, 1.25, 1.5$, or 2.0.
- **Rotation:** for each readout, we form a grayscale pixel-wise range map across the 60 MEDI frames. Pixels below 0.10 of the map maximum receive zero weight; the remaining single connected support is fit by a weighted, freely oriented ellipse. Its weighted centroid defines the rotation center, and its weighted second-moment eigenvectors define the ellipse orientation and axes. Every frame is rotated about that empirical center by

$\pm 2, \pm 5, \pm 10, \pm 15, \pm 30, \pm 60$, or $\pm 90^\circ$, using bicubic interpolation and frame-mean boundary filling followed by clipping to the model input range.

- **Translation:** the full sequence is shifted by 1, 2, 4, 8, 16, 32, or 48 pixels in each cardinal direction, with mean-value boundary filling.
- **Temporal reversal:** the complete frame sequence is presented in reverse order.
- **speed-up:** the sequence is played at 1.25, 1.5, or 2.0 times its nominal temporal speed, with $N_{\text{out}} = \lfloor (T_{\text{orig}} - 1)/s \rfloor + 1$ frames, where T_{orig} is the original frame count and s is the speed multiplier.
- **slow-down:** the sequence is played at 0.5, 0.67, or 0.8 times its nominal temporal speed, with $N_{\text{out}} = \lfloor (T_{\text{orig}} - 1)/s \rfloor + 1$ frames, where T_{orig} is the original frame count and s is the speed multiplier.

The parameter statistic is first averaged over repeated entries for each readout, preserving one value per readout and parameter. A condition-level summary, when used, is the median over the parameters listed for that transformation condition.

B.4 Statistical tests and dispersion

For each readout we first average repeated entries at a given transformation parameter, preserving one value per readout and parameter. Area differences are tested independently at each parameter with Kruskal–Wallis omnibus tests. Shaded bands in the main parameter plots are the standard deviation across readouts, not an uncertainty interval. Readout-bootstrap intervals and hierarchy-compatibility probabilities use 5,000 resamples throughout. We report parameter-specific correlations because a retention distribution concentrated near one has little available variation and can mechanically attenuate any correlation.

Selectivity differences among L2/3, L4, and L5 are tested separately within each visual area using Kruskal–Wallis tests. Two-sided Mann–Whitney tests then compare L2/3 with L4 and L5; Benjamini–Hochberg correction is applied separately to the four omnibus tests and the eight pairwise tests. The released `figures/tables/layer_selectivity_tests.csv` reports exact sample sizes, statistics, median differences, rank-biserial effect sizes, raw p values, and corrected q values. Readouts are model outputs rather than independent animals; these tests quantify consistency across model readouts, not biological replication. For the selectivity hierarchy check, we resampled readouts within each area 5,000 times and allowed either monotonic direction under the partial order V1→LM/RL→AL; neither direction occurred in any resample. The summary is in `figures/tables/hierarchy_consistency_selectivity.csv`.

B.5 Foundation-model evaluation

Each stimulus is converted to the spatial and intensity format expected by the foundation model. The model is frozen during all analyses. For the original 60-frame MEDI we average the predicted response over the final 50 frames after discarding the first 10 frames, which avoids treating recurrent warm-up as a stimulus effect. Variable-duration speed clips are evaluated over their complete generated sequence after the same 10-frame warm-up convention. A response ratio always uses the same readout in numerator and denominator. Retention is therefore a conditional model-output ratio, not a calibrated physiological fraction.

B.6 Data cohort and reporting scope

The complete functional manifest contains 104,171 readouts from four visual areas: V1 (75,497), LM (13,283), RL (11,232), and AL (4,159). These counts define the area-level analysis. A separate metadata table provides reliable cortical-layer labels for 38,381 readouts (L2/3: 20,898; L4: 11,703; L5: 5,780). Layer labels are incomplete because the functional and anatomical resources do not cover exactly the same units. We therefore use all readouts for area-level results and report estimates by cortical layer only where a label is present. Structural identifiers are used for metadata joins and reproducibility, not as a prerequisite for a functional readout to enter the area analysis.

B.7 COHORT WITH CORTICAL-LAYER LABELS

The layer-labelled total is 38,381 (L2/3: 20,898; L4: 11,703; L5: 5,780); the remainder has no reliable layer assignment. Table 1 tabulates the resulting area-by-layer availability, including the unassigned portion in every area.

Table 1: **Availability of cortical-layer labels.** Entries are readout counts. The three layer columns and their labelled total define the cohort used for comparisons by cortical layer; "Unassigned" is the remainder without a reliable layer label, and the last column is the complete area-level cohort. All readouts remain in area-level analyses.

Area	L2/3	L4	L5	Labelled total	Unassigned	Full area total
V1	13,627	8,072	3,347	25,046	50,451	75,497
LM	161	93	30	284	12,999	13,283
RL	5,437	2,666	1,736	9,839	1,393	11,232
AL	1,673	872	667	3,212	947	4,159
Total	20,898	11,703	5,780	38,381	65,790	104,171

B.8 SOURCE-LAYER BY TARGET-LAYER RESPONSE BLOCKS

The full response matrix also permits a direct block analysis by cortical layer. For each completed source/target block, rows were selected by the layer label attached to the source MEDI and columns by the layer label attached to the target readout. Rebalanced copies were deduplicated within each target block using the source task index, while the same source was retained across different target blocks because those blocks contain different target columns. The resulting block means are: L2/3→L2/3, 0.855; L2/3→L4, 0.827; L2/3→L5, 0.806; L4→L2/3, 0.829; L4→L4, 0.812; L4→L5, 0.783; L5→L2/3, 0.852; L5→L4, 0.834; and L5→L5, 0.821. These are mean predicted responses, not normalized connectivity strengths or a released connectivity table.

C QUALITY CONTROLS AND AREAL TOLERANCE ANALYSES

C.1 PARAMETER-RESOLVED AREAL TOLERANCE PROFILES

The parameter-resolved profiles are shown in Figure 11. Each panel gives median retention for every tested parameter, separately for V1, LM, RL, and AL; markers and connecting lines show area medians and translucent bands show one standard deviation across readouts. Horizontal axes retain their native parameter units, and speed-up and slow-down remain separate conditions.

C.2 QUALITY CONTROLS

Quality filters are defined globally before area summaries: "original response" is the unmodified-MEDI denominator, and $cc_{\max}$ is the maximum Pearson correlation between predicted and held-out measured response traces. The q05+, q10+, and q25+ filters retain readouts above the global 5th, 10th, and 25th response percentiles; cc .3+, .5+, and .7+ require $cc_{\max} \geq 0.3, 0.5, 0.7$. All retained counts and contrasts are reported separately for each area in the accompanying CSV file. Area contrasts are recomputed after filtering readouts by the model's prediction ceiling. Figure 12 reports individual-readout quality sensitivity separately for contrast, rotation, translation, temporal reverse, speed-up, and slow-down. Under the prespecified stability rule, an areal contrast is stable only when every displayed quality-threshold interval excludes zero. Translation satisfies this rule; rotation and temporal reversal do not, because at least one threshold interval crosses zero. These checks address sensitivity to readout quality and do not constitute independent biological validation.

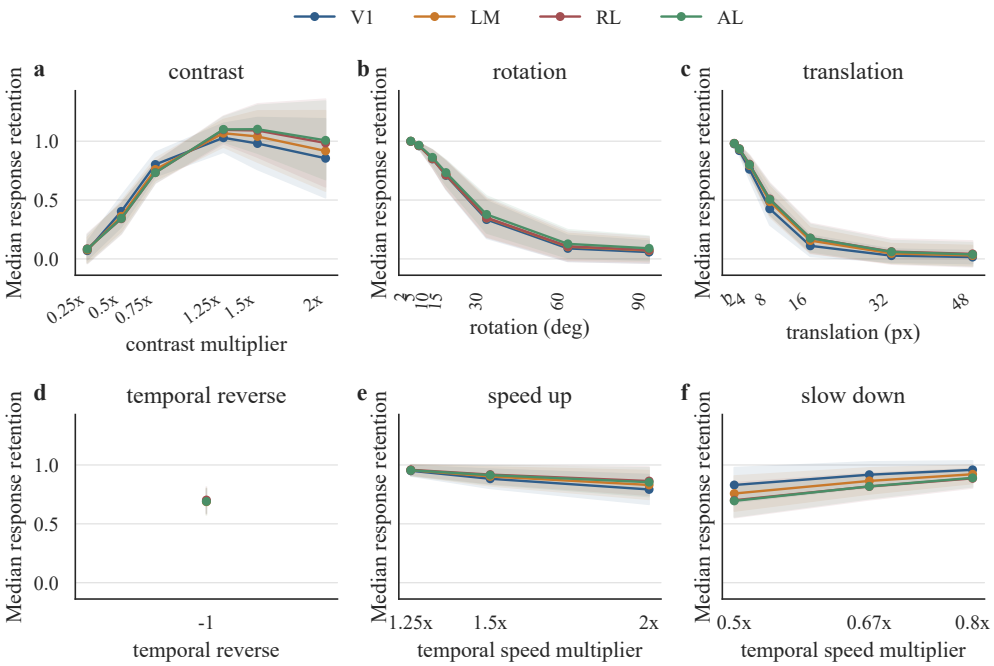


Figure 11: **Parameter-resolved areal retention profiles.** The six panels correspond to contrast, rotation, translation, temporal reverse, speed-up, and slow-down. Colored lines and markers identify areas; translucent bands show one standard deviation across readouts. Horizontal axes use native parameter units, including the reversal code -1.

D Supplementary selectivity–tolerance analyses

D.1 Self-excluded selectivity–tolerance analysis

To test whether the target readout's own MEDI mechanically contributes to the selectivity–tolerance association, we recompute its area-balanced lifetime sparseness after removing that MEDI from the source-area pool used for that target. The leave-one-out score is formed from the same source-pool sums and sums of squares as the full score, with the diagonal response and one sample removed only from the target's own source-area pool; the other three pools are unchanged. Figure 13 repeats the selected-stress analysis in Figure 5 with this self-excluded selectivity score. The correlations are nearly unchanged, showing that the reported associations are not driven by the target's own MEDI entering both measurements. Numerical values are provided in `figures/tables/selectivity_tolerance_loo_by_parameter.csv`.

D.2 Parameter-resolved selectivity–tolerance correlations

The main selectivity–tolerance panel uses the lowest-median-retention parameter for each transformation so that the displayed correlations are evaluated where retention has appreciable variation. Figure 14 shows the complete parameter grid. We also report the complete parameter grid here: each point is the Spearman correlation between area-balanced MEDI-set stimulus selectivity and retention for one area and one tested parameter. Parameters with retention close to one can show attenuated correlations because the response-retention range is compressed; including the full grid prevents the single-parameter display from being mistaken for a claim that other tested values were uninformative.

D.3 Nested selectivity–tolerance analysis

The strongest perturbation for each condition is selected using only one deterministic half of the readouts. Spearman ρ is then evaluated in the other half, separately by area, and the

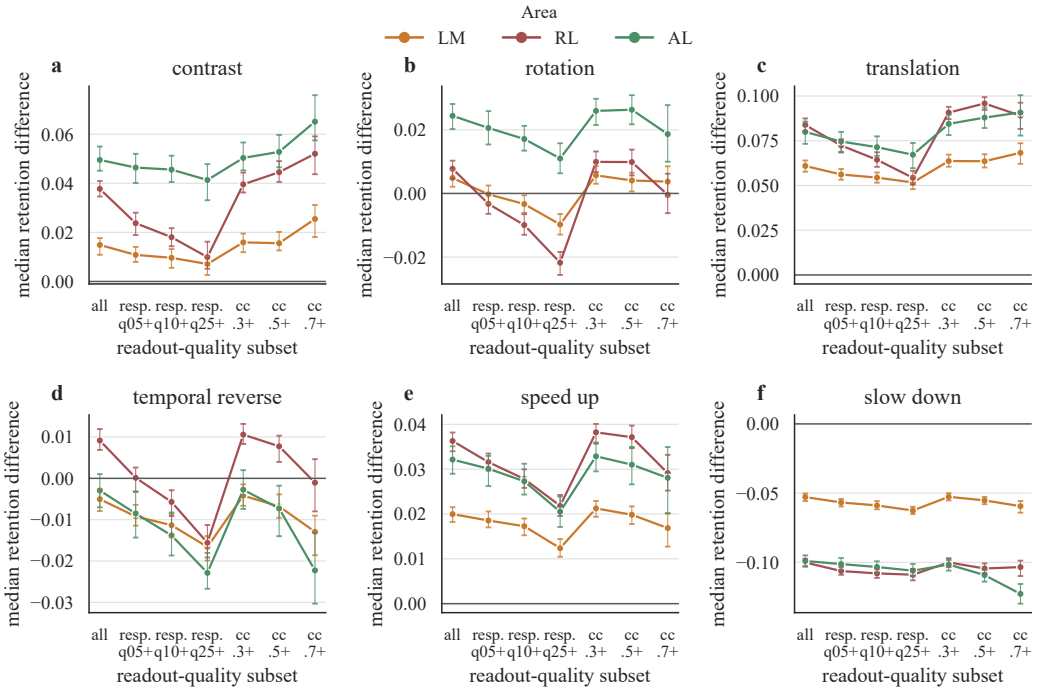


Figure 12: **Quality-threshold sensitivity.** The six panels show contrast, rotation, translation, temporal reverse, speed-up, and slow-down separately. The horizontal axis lists the unfiltered cohort, original-response quantile thresholds, and prediction-correlation thresholds; retained readout counts are area-specific and are provided for every panel and filter in `figures/tables/quality_sensitivity_summary.csv` (for example, unfiltered V1/LM/RL/AL = 75,497/13,283/11,232/4,159; q25+ = 59,888/9,100/6,450/2,691). Markers are median contrasts, thin vertical bars are 95% readout-bootstrap intervals, and connecting lines aid comparison across filters. Colors identify LM, RL, and AL contrasts relative to V1, and the gray horizontal line marks zero difference. Translation remains positive across all displayed thresholds; rotation and temporal reverse do not satisfy the all-intervals-exclude-zero stability rule. speed-up and slow-down are shown separately.

roles are exchanged. This preserves the intended stress-test interpretation while separating parameter selection from estimation. The selection parameters are contrast $0.25\times$, rotation $90°$, translation 48 pixels, temporal reverse, speed-up $2\times$, and slow-down $0.5\times$ in both directions of the split. Figure 15 shows the two held-out estimates and their bootstrap intervals. Numerical values, including FDR-adjusted p values and exact testing n, are tabulated in `figures/tables/nested_selectivity_tolerance_audit.csv`.

E Supplementary validation and descriptive analyses

E.1 Layer-by-transformation estimates

Layer tolerance is computed with the same ordinary response-retention ratio as in the area analysis, then summarized within each area–layer stratum. Figure 16 renders the median across tested parameters for every condition, with every cell's labelled sample size. V1 L4 has median translation retention 0.392 versus 0.436 in V1 L2/3, while median slow-down retention is 0.957 versus 0.906; in RL, L5 has median translation retention 0.546. These values are descriptive contrasts, not a universal laminar law. The complete 324-row parameter-by-layer table (four areas, three layers, six conditions, and all tested parameters) is supplied as `figures/tables/area_layer_transformation_summary.csv`; each row reports the median retention and the exact labelled sample size. LM L4 ($n = 93$) and LM L5 ($n = 30$) should not be overinterpreted.

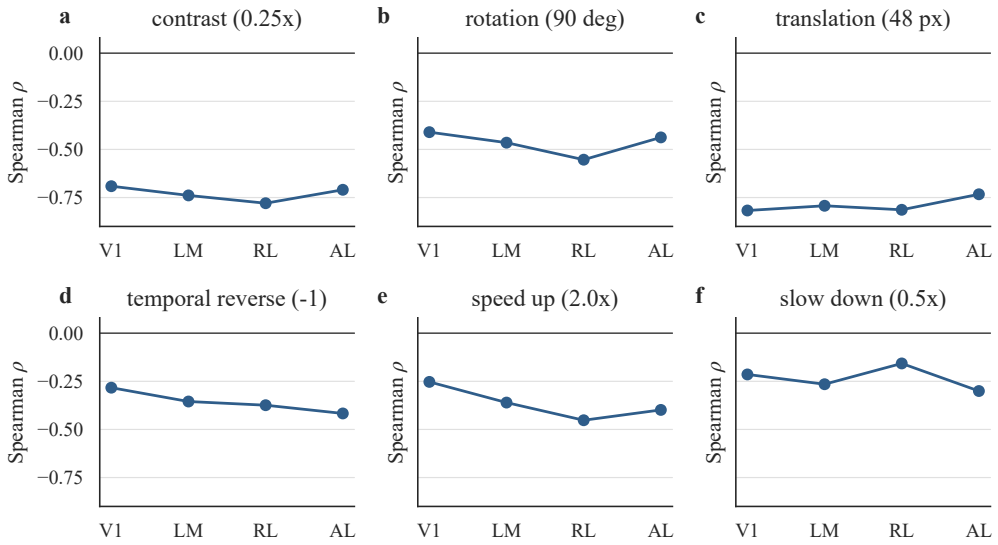


Figure 13: **Selected-stress selectivity–tolerance associations after excluding each target readout's own MEDI from its selectivity calculation.** The layout and selected parameters match Figure 5. Connected blue points show Spearman ρ for V1, LM, RL, and AL; the gray line marks zero.

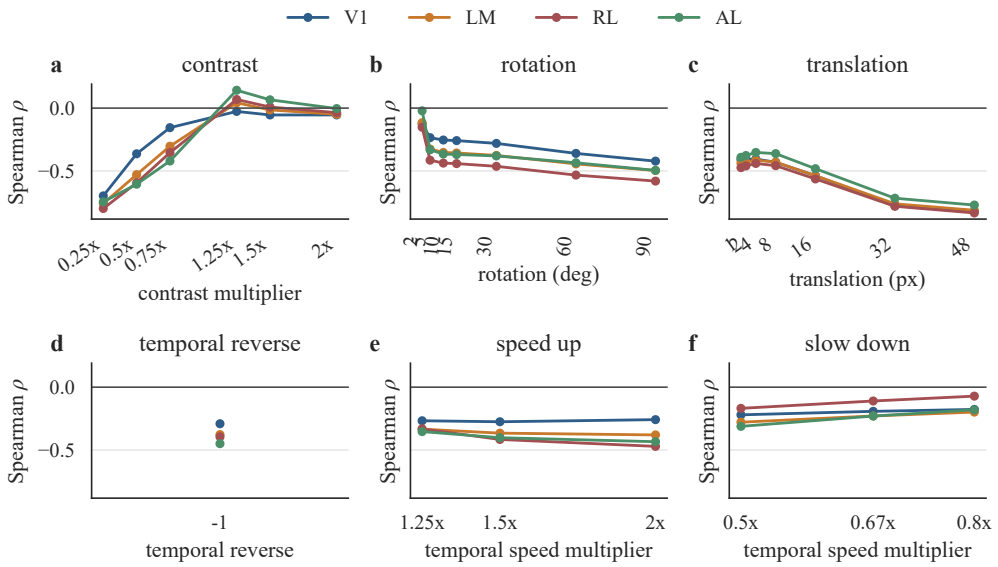


Figure 14: **Complete parameter-resolved selectivity–tolerance correlations.** The six panels correspond to contrast, rotation, translation, temporal reverse, speed-up, and slow-down. Within each panel, colored lines and markers show Spearman ρ between area-balanced MEDI-set stimulus selectivity and retention for V1, LM, RL, and AL across every tested parameter; the horizontal axis uses the parameter's native units, including multipliers, degrees, pixels, the reversal code -1, and speed multipliers. No uncertainty band is shown. The horizontal zero line is a visual guide. This full grid complements Figure 5, which selects the lowest-median-retention parameter for a compact, stress-focused comparison.

E.2 Temporal-reverse mechanism screen

We classify a readout as a near-static candidate when it lies in the top decile of reversal retention but below the within-area third quartile of forward/reverse pixel distance. A readout is classified as response-invariant despite large change when it lies in both the top retention decile and the top

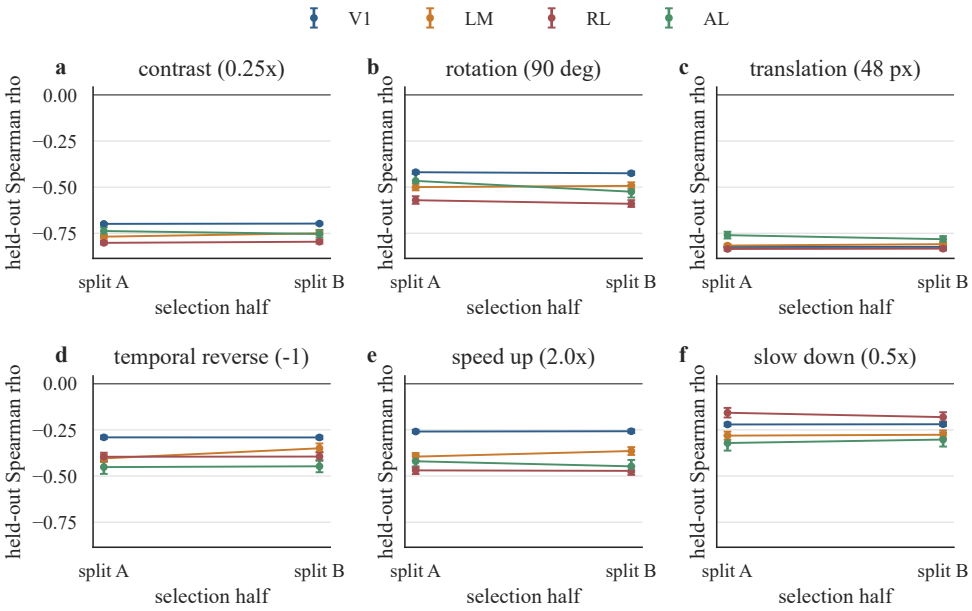


Figure 15: **Nested split analysis of selectivity–tolerance associations.** Lowercase panel titles give the transformation and selected parameter with its unit. Colors identify visual areas. At split A and split B, points show held-out Spearman ρ after the opposite half selected the lowest-median-retention parameter; connecting lines link the two exchanged selection/test roles. Vertical error bars are 95% percentile bootstrap intervals over readouts, and the gray horizontal line marks zero association. The analysis includes exact test samples and FDR-adjusted p values.

pixel-distance quartile. Figure 17 visualizes the three resulting groups within each area. The screen is descriptive: pixel distance measures stimulus change, not a neural mechanism. Counts and identifiers are in `figures/tables/temporal_reverse_two_mechanisms.csv`. Figure 18 provides representative MEDI frames for the two high-retention categories.

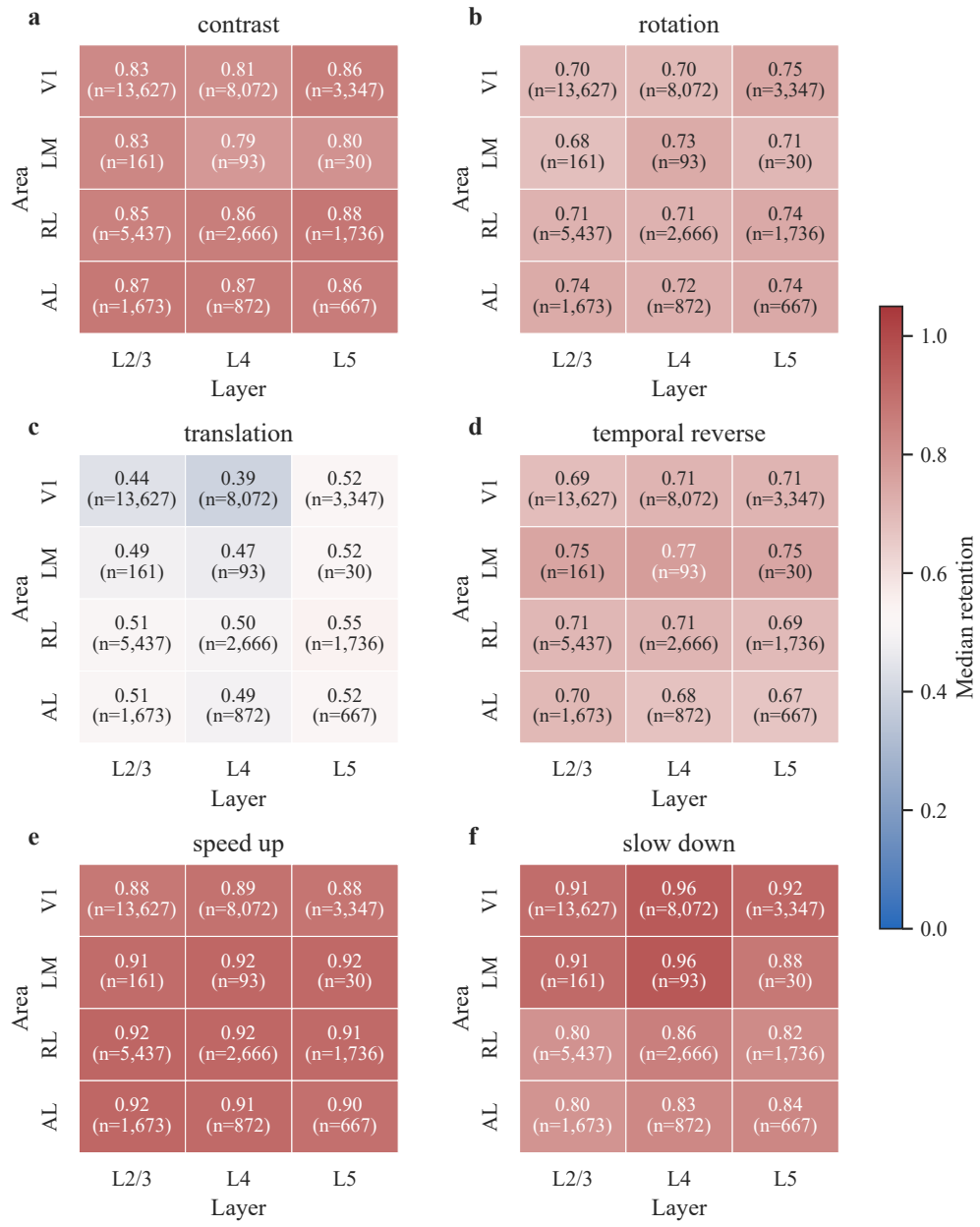


Figure 16: **Layer-by-condition retention summaries.** The six panels correspond to the labelled transformation conditions. Within each panel, rows are visual areas and columns are cortical layers; tile color and the first printed number encode median response retention on the shared color-bar scale, the parenthetical value is the exact labelled sample size, and white borders delimit tiles. This summary is descriptive; parameter-resolved values are supplied in `figures/tables/area_layer_transformation_summary.csv`. LM L4 and L5 have very small samples.

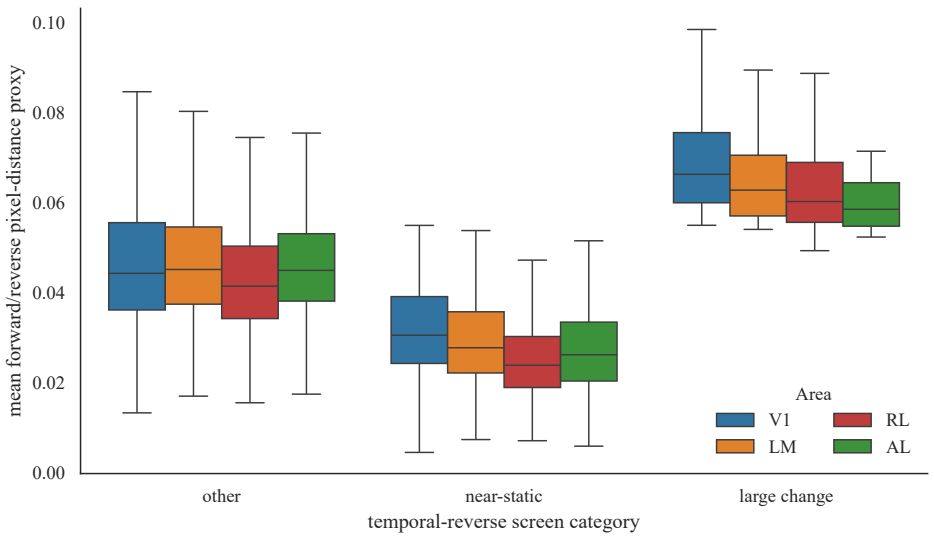


Figure 17: **Temporal-reverse mechanism screen.** The horizontal axis gives the three screening categories and colors identify visual areas. For each area and category, the box spans the interquartile range, the internal line is the median, and whiskers extend to the most extreme observation within 1.5 interquartile ranges; more distant points are omitted from display. The vertical axis is the forward/reverse pixel-distance proxy. The large-change category identifies high predicted retention that is inconsistent with a nearly static stimulus under the defined pixel-distance proxy.

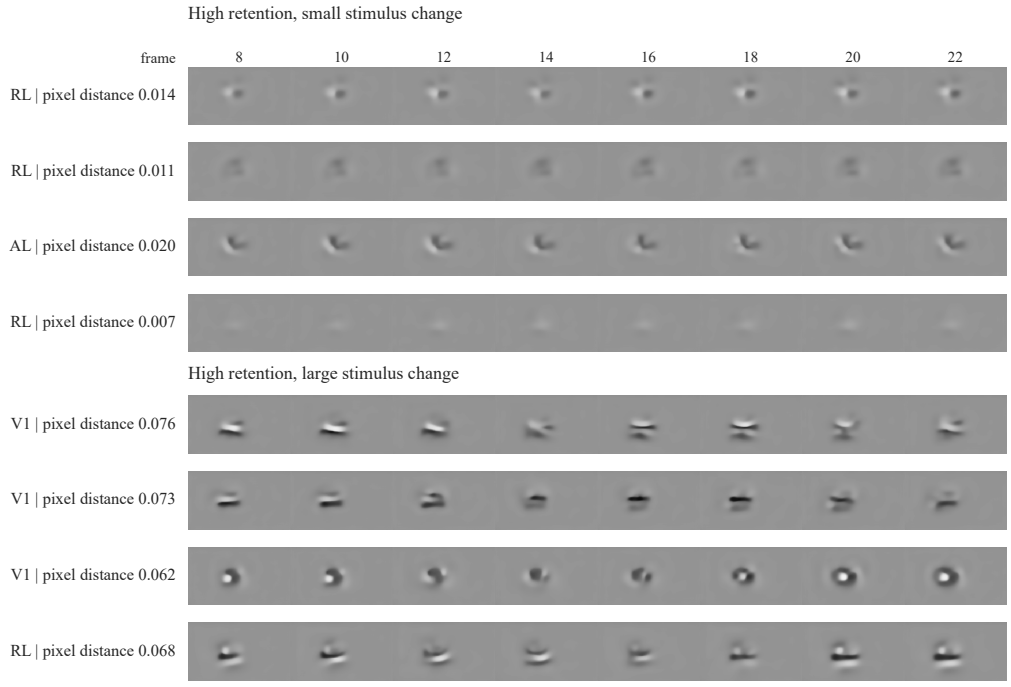


Figure 18: **Representative temporal-reverse mechanisms.** The two group headers identify high-retention MEDIs with small and large stimulus change. Each category contains four rows. The left-hand annotation on each row gives visual area and forward/reverse pixel distance; the frame labels above the image columns apply to every row. These examples illustrate the screening categories and are not independent statistical samples.