

Equilibrium Propagation Discovers Top-Down Feedback for Audio-Visual Binding in Continuous Wave Fields

Jeremy Slater

dr.slater@stratusneuro.com

MERLN LLC <https://orcid.org/0000-0002-2656-1354>

Gardar Thorvardsson

Kvikna Medical

Research Article

Keywords: neuromorphic computing, Landau-Ginzburg dynamics, equilibrium propagation, wave computing, continuous wave fields, multimodal integration

Posted Date: April 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9404804/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare potential competing interests as follows: J.D.S. is Chief Medical Officer of Stratus Neuro, Chief Innovation Officer of MERLN LLC, and founder of Medscrios LLC. G.T. is managing director of Kvikna Medical. The authors declare that none of these affiliations influenced the design, execution, or interpretation of the reported research.

Abstract

Cross-modal binding — the fusion of simultaneous sensory streams into a unified percept — has not been achieved in physical neural networks without backpropagation. Whether top-down feedback between hierarchical field layers can emerge from local learning rules alone remains untested. We extend a Landau-Ginzburg wave field architecture trained by Equilibrium Propagation to a two-layer system: primary audio and visual fields drive a binding field that sends top-down feedback to both primaries through coupling coefficients initialized to zero. Trained on the GRID audiovisual corpus, the coupling coefficients grow from 0.0 to 0.051 over ten epochs — a result absent in the unimodal case — confirming that Equilibrium Propagation discovers top-down feedback when cross-modal binding is required. The binding field outperforms late fusion; replacing phase-sensitive measurement with amplitude-only readout costs 9.2 percentage points, exceeding the analogous unimodal penalty. When presented with conflicting audiovisual inputs, the system produces fusion responses in 83% of trials, stable under contrastive readout training and therefore reflecting field dynamics rather than readout bias. Symmetric noise degradation — 33.3 versus 33.7 percentage points for audio and video respectively — confirms genuine integration.

Introduction

The brain fuses what it sees and hears into a single percept so seamlessly that the seams only become visible when the modalities conflict. When auditory /ba/ is dubbed onto visual /ga/, listeners report hearing /da/ — a syllable present in neither stream¹. The McGurk effect is not a laboratory oddity. It demonstrates that the auditory percept is not simply what arrives at the ear; it is constructed from both streams weighted by their reliability, and the weighting happens automatically, below the level of conscious control. How this works has been worked out in some detail. Audiovisual binding operates within a temporal window of roughly 200 milliseconds², and neuroimaging has shown that the computation is distributed hierarchically across cortex: primary sensory areas process each modality independently, a parietal stage fuses them under the default assumption of a common source, and an anterior stage performs the full causal inference — asking whether the signals actually belong together before committing to a fused percept³.

What makes this hierarchy function is top-down feedback. Feedforward signals propagate up the hierarchy in the gamma band (30–70 Hz); predictions flow back down in alpha and beta (8–30 Hz)^{4,5}. The higher areas generate expectations of what the lower areas are about to receive and suppress the response when those expectations are met — passing forward only the residual, the part that was not predicted⁶. When this feedback is pharmacologically disrupted, as during propofol-mediated loss of consciousness, the hierarchy collapses: predicted stimuli are no longer suppressed, unpredicted stimuli are no longer selectively amplified, and the binding that depended on the interplay between the two directions of signaling disappears⁷. For any physical computing system attempting genuine cross-modal

integration, this is the essential point: top-down feedback between hierarchical layers is not a refinement that can be added later — it is what binding is.

Existing computational approaches to audiovisual speech integration have set this aside by necessity. Backpropagation-trained architectures — late fusion systems that combine learned unimodal representations at the classifier, early fusion systems that merge raw features before encoding, and attention-based models that learn cross-modal correspondences through transformers^{8,9} — achieve strong performance, but the binding they implement is a learned weight matrix rather than a physical process. This matters for hardware. In a physical neural network, the forward computation occurs in the substrate itself and no global backward pass is available. The one multimodal physical system demonstrated to date — a trainable diffractive optical chip that classifies visual, auditory, and tactile stimuli — achieves 85.7% accuracy but trains through backpropagation applied to a digital twin¹⁰. No physical neural network has learned cross-modal binding through local rules alone, and none has shown top-down feedback emerging from learning rather than being hardwired in.

In a companion paper we showed that a continuous Landau-Ginzburg (LG) wave field trained by Equilibrium Propagation (EP) achieves 74.1% accuracy on Google Speech Commands V2 with no backpropagation through the field dynamics¹¹. EP updates physical parameters using only the difference in local field statistics between a freely-settled state and a weakly nudged one — a purely local computation¹². One result from that work bears directly on what follows: EP drove the top-down feedback coefficient λ_{td} to zero in the unimodal setting. A single sensory stream offered the field no reason to develop predictions of its own inputs. The prediction, then, is that λ_{td} should grow when the field is presented with two streams that need to be bound — when there is something worth predicting across the modality boundary.

We test this by extending the architecture to two primary fields (L1 Audio and L1 Visual) plus a binding field (L2) that receives bottom-up drive from both and can send top-down feedback to each through λ_{td} values initialized at zero. The system is trained on the GRID audiovisual sentence corpus¹³, 34 speakers producing 1,000 sentences each in a controlled vocabulary, with aligned audio and lip-region video available per word token. EP discovers top-down feedback: λ_{td} grows from 0.0 to 0.051 over 10 training epochs, whereas it remained at zero in the unimodal case. The L2 field beats late fusion by 5.2 percentage points, confirming that the wave field dynamics contribute binding beyond feature concatenation. Replacing the phase-sensitive readout with amplitude-only measurement costs 9.2 percentage points — more than the 7.8-point cost in the unimodal L1 field, suggesting that explicit phase extraction becomes more important as the representation must carry cross-modal rather than single-stream information. Finally, degrading audio to 0 dB SNR and degrading video to 0 dB SNR produce nearly symmetric accuracy drops (33.3 and 33.7 points respectively), indicating the two modalities are genuinely integrated rather than one quietly dominating.

Results

Equilibrium Propagation discovers top-down feedback in the multimodal setting

The L2 binding field was initialized with top-down coupling coefficients $\lambda_{td_audio} = \lambda_{td_visual} = 0.0$, placing both fields in a regime where the binding field receives sensory drive from both L1 fields but sends nothing back. Over ten EP training epochs, both coefficients grew monotonically to 0.051, at a rate of approximately 0.005 per epoch (Fig. 1a). In the companion unimodal system, EP drove λ_{td} to zero and held it there¹¹. The learning rule, initialization, and EP hyperparameters are identical across the two settings; the presence of a second sensory stream is the only structural difference.

The physics parameters evolved alongside λ_{td} . Damping rates decreased uniformly across all three channels (γ : 0.101 $\rightarrow$ 0.096, - 4.7%), and lateral inhibition strength decreased equally (D : 0.261 $\rightarrow$ 0.250, - 4.4%) (Fig. 4a,b). Both changes are in the same direction as the L1 audio field in the companion paper — lower dissipation, longer field memory, tighter spatial coupling — suggesting a general principle: EP drives LG fields toward low-dissipation regimes when the task requires preserving phase structure. The saturation coefficient β and rotation frequency ω were not updated by EP in the L2 training loop and remained at initialized values throughout.

Validation accuracy grew from 32.8% at EP epoch 0 to 42.8% by epoch 9, then to 45.6% during 50 epochs of readout optimization with frozen physics (Fig. 1b). The late fusion baseline — concatenating the same L1 audio and visual features without L2 field dynamics — reached 40.0%. The L2 field adds 5.6 percentage points over feature concatenation alone.

Phase-sensitive readout is more critical in the multimodal than the unimodal setting

Systematic ablation across ten conditions reveals the relative contribution of each architectural component (Table 1, Fig. 2). Replacing the phase-sensitive six-feature readout with amplitude-only measurement (ablation 5) costs 9.2 percentage points — larger than the 7.8-point cost in the unimodal L1 field. The binding field must represent not just the content of each modality but their cross-modal relationship, which is carried in the relative phase structure of the two L1 field states; an amplitude-only readout cannot recover that relationship.

Table 1
Ablation study results

#	Condition	Val. Acc. (%)	Δ (pp)
0	Full L2 System (control)	45.6	-
1	Audio-only L1	44.3	-1.4
2	Visual-only L1	24.9	-20.8
3	Late fusion (no L2 field)	40.0	-5.6
4	L2 without top-down feedback ($\lambda_{td} = 0$)	45.6	-0.1
5	Amplitude-only readout	36.4	-9.2
6	L2 initialized physics (no EP)	45.8	+ 0.2
7	McGurk matched pairs	45.6	0.0
8	Noisy audio (0 dB SNR)	12.4	-33.3
9	Noisy video (0 dB SNR)	12.0	-33.7
10	Phase normalization	37.0	-8.6
11	+Contrastive loss	46.2	+ 0.6
<p>Each row shows the effect of removing or modifying one component of the L2 binding system. Physics parameters are frozen at trained values for all ablations except ablation 6 (physics reset to initialization). Δ indicates accuracy change from the full L2 system control (ablation 0, 45.6% validation accuracy). All accuracies are known-class accuracy computed over the 51 GRID word classes on the validation split (N = 12,000 samples, speakers 29–32). Ablation 11 retraines the readout with an auxiliary proxy-based contrastive loss (see Methods); mismatch rejection rate is the fraction of conflicting AV pairs for which the predicted class matches neither the audio class label nor the video class label.</p>			

Removing top-down feedback entirely (ablation 4, λ_{td} frozen at 0) costs 0.1 percentage points, and resetting L2 physics to initialization without EP (ablation 6) improves accuracy by 0.2 points — both within noise. This mirrors the pattern in the unimodal paper, where individual physics components each contributed under one percentage point. The field dynamics and top-down feedback are not what drives the headline accuracy number; they drive the binding behavior described below.

The modality asymmetry is stark. Visual-only input (ablation 2) produces 24.9% — substantially above chance (2.0%) but 20.8 points below the full system, reflecting the inherent difficulty of lip reading relative to audio classification on GRID. Audio-only input (ablation 1) produces 44.3%, 1.4 points below full L2. The full system at 45.6% exceeds audio-only despite the visual stream being the weaker modality by a wide margin, confirming the visual stream contributes discriminative information to the binding field rather than diluting it.

Noise robustness reveals symmetric cross-modal integration

Adding Gaussian noise to the audio stream at 0 dB SNR (ablation 8) drops accuracy from 45.6% to 12.4% (– 33.3 pp). Adding equivalent noise to the video stream (ablation 9) drops accuracy to 12.0% (– 33.7 pp) (Fig. 3). The two drops are separated by 0.4 percentage points – smaller than the measurement noise across individual ablations. In a system where one modality dominates, degrading the dominant stream would cost far more than degrading the weaker one. The near-identical degradation shows the L2 field distributes its reliance across both streams in roughly equal measure.

Conflicting audiovisual inputs produce fusion responses

Conflicting audiovisual inputs produce fusion responses

When presented with 4,320 conflicting audiovisual pairs – audio from one word class dubbed onto video from a different class, constructed using viseme-based pairing to maximize perceptual conflict – the L2 system produced fusion responses in 83.0% of trials. Audio capture occurred in 6.9% of trials and visual capture in 10.1% (Fig. 5a). The late fusion baseline produced audio capture in 20.9% of trials and fusion in 77.2% – the full L2 system shifts resolution toward fusion by 5.8 percentage points.

To determine whether the 83% fusion rate reflects the readout layer or the underlying field dynamics, we retrained the readout with a proxy-based contrastive loss designed to push mismatched AV representations away from all class proxies (ablation 11). Classification accuracy improved by 0.6 percentage points (45.6% → 46.2%) and the fusion rate held at 83.2% (Fig. 5b). The resolution strategy survived explicit pressure to change it, placing the 83% fusion rate in the field dynamics rather than the readout geometry.

Phase coherence between the frozen L1 audio and visual fields – measured as the mean complex exponential of the inter-field phase difference across spatial locations and channels – held at 0.043 throughout all 60 training epochs. Relative phase normalization (ablation 10) raised the raw coherence value to 0.343 but produced no matched-versus-mismatched coherence gap. Under current training, the binding information resides in the amplitude-domain features of the field states rather than their phase relationship.

The speaker-independent test split (speakers 33–34, N = 6,000 samples) yielded 42.3% test accuracy against 45.6% validation, a 3.3-point gap attributable to the limited speaker diversity in the two held-out speakers. All other reported accuracies use the validation split (speakers 29–32) following the standard GRID evaluation protocol.

Discussion

The L2 binding field was initialized with $\lambda_{td} = 0.0$ and left free to grow or remain there. It grew – steadily, linearly, over ten EP epochs – and did not grow in the unimodal system running the same

learning rule on the same hardware. The structural difference between the two settings is the presence of a second sensory stream. In biological multisensory cortex, feedforward projections operate in the gamma band and feedback projections in alpha and beta^{4,5} — the same directional asymmetry, bottom-up drive and top-down binding signal as structurally separate channels, emerges in both biological systems and in EP-trained LG fields from different starting points and by different mechanisms. Whether this reflects something deep about the mathematics of binding or is a surface coincidence, photonic hardware implementation could help answer: in a substrate where field frequencies are directly measurable, the predicted spectral signature of λ_{td} feedback would be testable in ways that simulation cannot provide.

The 83% fusion rate on conflicting audiovisual pairs fits the causal inference account of multisensory perception¹⁴. When two streams conflict sufficiently that neither source hypothesis dominates, the causal inference model predicts fusion rather than selection — and that is what the L2 field produces. More tellingly, it produced this before contrastive training and continued producing it after, with the fusion rate shifting by only 0.2 percentage points despite the readout being explicitly retrained to push mismatched representations apart. The resolution strategy is in the field, not the readout. Late fusion, with no binding field, produced audio capture in 20.9% of trials — the stronger audio stream dominated when no intermediary mediated the conflict.

Phase coherence between the frozen L1 fields held flat at 0.043 for all 60 training epochs. Relative phase normalization raised the raw value to 0.343 but produced no gap between matched and mismatched pairs. The binding information under current training sits in amplitude-domain features rather than phase relationships. In the biological predictive coding literature, phase coherence as a binding signal is associated with architectures where top-down predictions pre-activate lower areas before input arrives, registering binding as alignment between predicted and actual field states⁶. Standard EP contrasts free-phase and nudged-phase statistics — it does not generate predictions ahead of input. A training regime in which physics updates occur at moments of cross-modal prediction failure rather than continuously across all samples may be what is needed to induce phase coherence as a functional binding signal; the current architecture is already instrumented for that experiment.

The modality gap on GRID — audio at 47.7% versus visual at 26.1% — means the system operates throughout in a regime of pronounced unimodal asymmetry, and the + 5.6 pp advantage of L2 over late fusion may understate what is available when both streams contribute comparable information. The speaker-independent test split uses only two speakers, producing a 3.3-point validation-to-test gap that reflects sampling variance rather than a systematic failure. All results reflect single training runs; the small differences between ablation conditions — particularly the top-down and EP ablations at 0.1 and 0.2 points respectively — should be read in that light.

The lateral inhibition strength D , the damping rate γ , and the inter-layer coupling coefficient λ_{td} each map to a physical fabrication specification — a waveguide geometry, a material absorption coefficient, a coupling gap. The EP-discovered λ_{td} value of 0.051 is not just a learned hyperparameter; it is a target

coupling strength that could in principle be built into photonic hardware directly, transferring the binding geometry from simulation to substrate without retraining.

Methods

Dataset and evaluation protocol

The GRID audiovisual sentence corpus¹³ consists of 1,000 sentences spoken by each of 34 talkers (18 male, 16 female), recorded under controlled conditions with synchronized audio and frontal-view video. Sentences follow a fixed grammatical template ("command colour preposition letter digit adverb"), yielding a closed vocabulary of 51 word types. Audio is provided at 25 kHz; video as JPEG frames at 25 fps. Speaker 21 has no video and was excluded. We adopted the standard speaker-independent split: speakers 1–28 for training (66,000 word samples), 29–32 for validation (12,000), and 33–34 for test (6,000). Word-level boundaries were extracted from the provided forced-alignment files. All accuracy figures in the main text refer to the validation split unless otherwise noted.

Audio and visual front-ends

Audio segments were resampled to 16 kHz, zero-padded or trimmed to one second centred on the word boundary, and processed through the same mel spectrogram front-end as the companion paper: 400-sample Hann window, 160-sample hop, 64 mel bands, log compression, per-sample z-score normalization, with first and second temporal derivatives as channels 1 and 2. The resulting $3 \times 101 \times 64$ tensor was bilinearly interpolated to 94×64 to match the L1 Audio field resolution, time on the x-axis, frequency on the y-axis.

For visual input, the lip region of interest was extracted from each video frame using MediaPipe Face Mesh¹⁵ landmarks. The 20 lip-specific landmarks defined a bounding box with 20-pixel padding, cropped and resized to 64×32 pixels per frame. Word segments were resampled to 10 frames per word. A spatiotemporal representation was constructed by stacking frames along the temporal axis to yield a $64 \times 32 \times 10$ volume, then applying the same three-channel delta/delta-delta construction as the audio front-end and resizing to 94×64 .

L1 field architecture

The L1 Audio field is the trained system from the companion paper, loaded from checkpoint and frozen throughout all L2 experiments. It consists of a $94 \times 64 \times 3$ complex-valued Landau-Ginzburg field with phase-sensitive readout producing 162-dimensional feature vectors (6 features $\times$ 3 channels $\times$ 3×3 spatial pooling). The L1 Visual field uses an identical architecture trained from scratch on GRID visual word tokens: 20 epochs of joint physics-and-readout training followed by 50 epochs of readout optimization with frozen physics. Both L1 fields were frozen before L2 training.

L2 binding field and Equilibrium Propagation

The L2 binding field is a $47 \times 32 \times 3$ complex-valued LG field (half the L1 spatial resolution) initialized in the underdamped regime with $\gamma = [0.025, 0.050, 0.100]$ per channel and $D = 0.100$. The field receives bottom-up drive from both frozen L1 fields: the 162-dimensional feature vectors from each L1 field are projected into L2 field space via learned 3×3 coupling matrices initialized to $0.1 \times \text{identity}$. Top-down feedback is applied to both L1 fields via scalar coupling coefficients $\lambda_{\text{td_audio}}$ and $\lambda_{\text{td_visual}}$, both initialized to zero. L2 settles for 60 timesteps per input ($dt = 0.07$, semi-implicit Euler with implicit damping).

EP updates three L2 parameter sets per training step: γ (via nudged-minus-free contrast in mean field amplitude), D (via nudged-minus-free contrast in lateral inhibition drive), and λ_{td} (via nudged-minus-free contrast in the L2 field response to top-down injection). The update rules are local, following the formulation in the companion paper. EP learning rate was 0.01; parameters were clamped to physically meaningful ranges ($\gamma \in [0.01, 0.5]$, $D \in [0.001, 1.0]$, $\lambda_{\text{td}} \in [0.0, 1.0]$). EP ran for 10 epochs followed by 50 epochs of readout-only optimization with physics frozen.

Ablation methodology

Each ablation loads the best L2 checkpoint, modifies one component, reinitializes the readout to zero, and retrains it for 50 epochs with L2 physics frozen. The contrastive loss ablation (ablation 11) additionally trains a ProxyContrastiveLoss module — a learned projection from 162-dimensional features to a 128-dimensional proxy space, with temperature $\tau = 0.07$ — alongside the standard delta-rule readout, using batches constructed with 75% matched and 25% mismatched pairs. The total loss is $L_{\text{cls}} + 0.1 \times L_{\text{contrastive}}$.

McGurk test construction

Conflicting audiovisual pairs were constructed by pairing audio tokens from one word class with video tokens from a visually similar class. Visual similarity was defined by cosine similarity between mean L1 visual feature vectors, computed over the training set. For each of the 51 word classes, the three most visually similar alternative classes served as hard negatives. The test set comprised 4,320 conflicting pairs and 1,020 matched pairs drawn from the validation split.

Computational resources

All experiments were conducted on a single NVIDIA DGX Spark GPU. L1 Visual training required approximately 12 GPU-hours; L2 EP training approximately 10 GPU-hours; each ablation approximately 2 GPU-hours for feature precomputation and readout retraining. Total: approximately 48 GPU-hours.

Declarations

Data availability

The GRID audiovisual sentence corpus is freely available for research use at <https://zenodo.org/record/3625687> (DOI: 10.5281/zenodo.3625687). The L1 Audio field checkpoint used in all L2 experiments is the trained system reported in the companion paper (ref. 12), which will be deposited in a public repository upon that paper's acceptance. L2 field checkpoints and ablation results will be deposited in the same repository upon acceptance of the present paper.

Code availability

Complete source code for the L2 binding field architecture, training pipeline, ablation framework, and McGurk test construction will be made available at <https://github.com/photodoc1960/ficu-l2> upon acceptance. The L1 Audio codebase from the companion paper will be made available concurrently.

Author contributions

J.D.S. conceived the architecture, designed the experiments, performed all computational work, analysed the data, and wrote the manuscript. G.T. provided critical review of the experimental design and manuscript.

Competing interests

J.D.S. is Chief Medical Officer of Stratus Neuro, Chief Innovation Officer of MERLN LLC, and founder of Medscrios LLC. G.T. is managing director of Kvikna Medical. The authors declare that none of these affiliations influenced the design, execution, or interpretation of the reported research.

Acknowledgements

The authors thank Giri Kalamangalam (University of Florida) for mathematical discussions which initially inspired the theoretical framework. Computing resources were provided by the DGX Spark platform. The GRID corpus was collected by Cooke, Barker, Cunningham and Shao at the University of Sheffield and made publicly available via Zenodo.

Use of AI

Artificial intelligence tools (Claude, Anthropic) were used to assist with computational simulation development, figure preparation, and manuscript writing. All AI-generated content was reviewed, validated, and revised by the authors, who take full responsibility for the accuracy and integrity of the work.

References

1. Tiippana K (2014) What is the McGurk effect? *Front Psychol* 5:725.
<https://doi.org/10.3389/fpsyg.2014.00725>
2. van Wassenhove V, Grant KW, Poeppel D (2007) Temporal window of integration in auditory-visual speech perception. *Neuropsychologia* 45:598–607.
<https://doi.org/10.1016/j.neuropsychologia.2006.01.001>
3. Rohe T, Noppeney U (2015) Cortical hierarchies perform Bayesian causal inference in multisensory perception. *PLoS Biol* 13:e1002073. <https://doi.org/10.1371/journal.pbio.1002073>
4. Michalareas G et al (2016) Alpha-Beta and Gamma Rhythms Subserve Feedback and Feedforward Influences among Human Visual Cortical Areas. *Neuron* 89:384–397.
<https://doi.org/10.1016/j.neuron.2015.12.018>
5. Bastos AM, Lundqvist M, Waite AS, Kopell N, Miller EK (2020) Layer and rhythm specificity for predictive routing. *Proc Natl Acad Sci USA* 117:31459–31469.
<https://doi.org/10.1073/pnas.2014868117>
6. Talsma D (2015) Predictive coding and multisensory integration: an attentional account of the multisensory mind. *Front Integr Neurosci* 9:19. <https://doi.org/10.3389/fnint.2015.00019>
7. Xiong YS et al (2024) Propofol-mediated loss of consciousness disrupts predictive routing and local field phase modulation of neural activity. *Proc Natl Acad Sci USA* 121:e2315160121.
<https://doi.org/10.1073/pnas.2315160121>
8. Ivanko D, Ryumin D, Karpov AA (2023) Review of Recent Advances on Deep Learning Methods for Audio-Visual Speech Recognition. *Mathematics* 11:2665
9. Michelsanti D et al (2021) An Overview of Deep-Learning-Based Audio-Visual Speech Enhancement and Separation. *IEEE/ACM Trans Audio Speech Lang Process* 29:1368–1396.
<https://doi.org/10.1109/TASLP.2021.3066303>
10. Cheng J et al (2024) Multimodal deep learning using on-chip diffractive optics with in situ training capability. *Nat Commun* 15:6189. <https://doi.org/10.1038/s41467-024-50677-3>
11. Slater J, Thorvardsson G (2026) Phase structure in continuous wave fields enables speech classification without backpropagation. Preprint Res Square. <https://doi.org/https://doi.org/10.21203/rs.3.rs-9205518/v1>
12. Scellier B, Bengio Y (2017) Equilibrium Propagation: Bridging the Gap between Energy-Based Models and Backpropagation. *Front Comput Neurosci* 11:24. <https://doi.org/10.3389/fncom.2017.00024>
13. Cooke M, Barker J, Cunningham S, Shao X (2006) An audio-visual corpus for speech perception and automatic speech recognition. *J Acoust Soc Am* 120:2421–2424.
<https://doi.org/10.1121/1.2229005>
14. Magnotti JF, Beauchamp MS (2017) A Causal Inference Model Explains Perception of the McGurk Effect and Other Incongruent Audiovisual Speech. *PLoS Comput Biol* 13:e1005229.
<https://doi.org/10.1371/journal.pcbi.1005229>
15. Lugaresi C et al (2019) MediaPipe: A Framework for Building Perception Pipelines. *arXiv [cs.DC]*
<https://doi.org/https://doi.org/10.48550/arXiv.1906.08172>

Figures

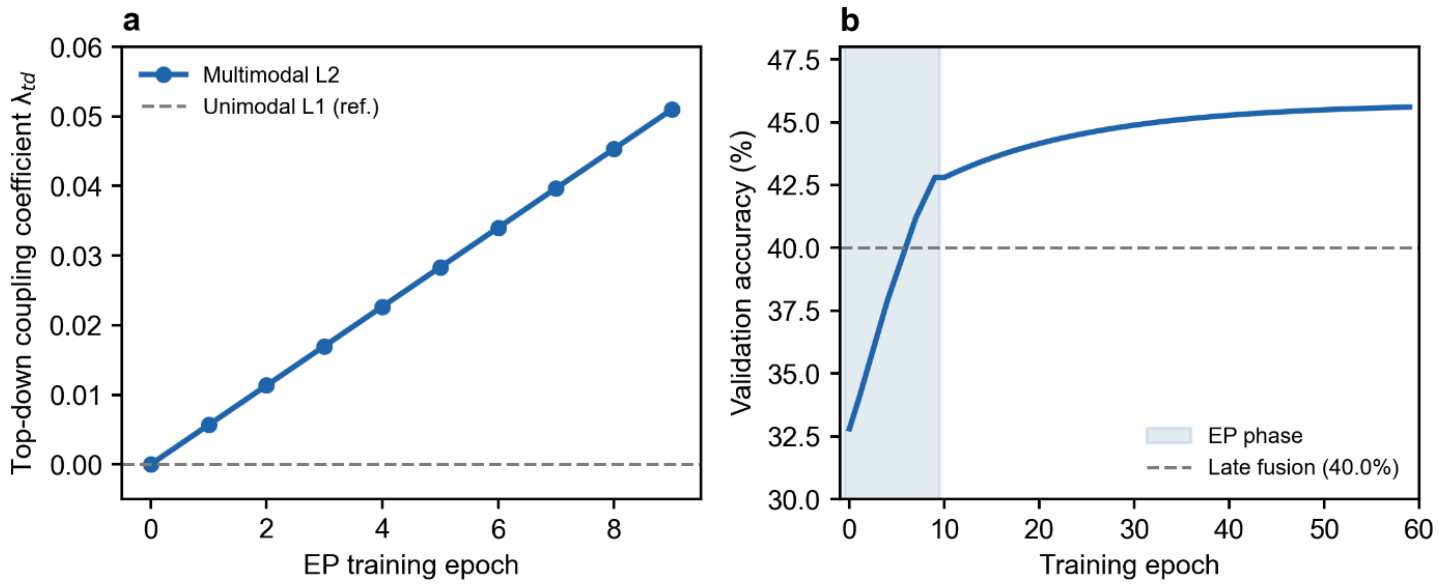


Figure 1

Equilibrium Propagation discovers top-down feedback in the multimodal setting.

a, Top-down coupling coefficients λ_{td} (audio and visual channels, values identical throughout training) as a function of EP training epoch. Both coefficients were initialized to zero and grew monotonically at approximately 0.005 per epoch, reaching 0.051 by epoch 9. The dashed line indicates the result from the companion unimodal system, where EP drove λ_{td} to zero and held it there across 100 training epochs. The learning rule, initialization, and hyperparameters are identical across the two settings; the presence of a second sensory stream requiring binding is the only structural difference. **b**, Validation accuracy over the full training arc. Blue shading indicates the EP phase (epochs 0–9), during which L2 physics and λ_{td} are jointly updated; white indicates the readout optimization phase (epochs 10–59), during which physics are frozen. The dashed horizontal line shows the late fusion baseline (40.0%), which concatenates the same L1 audio and visual features without L2 field dynamics. The full L2 system reaches 45.6% validation accuracy.

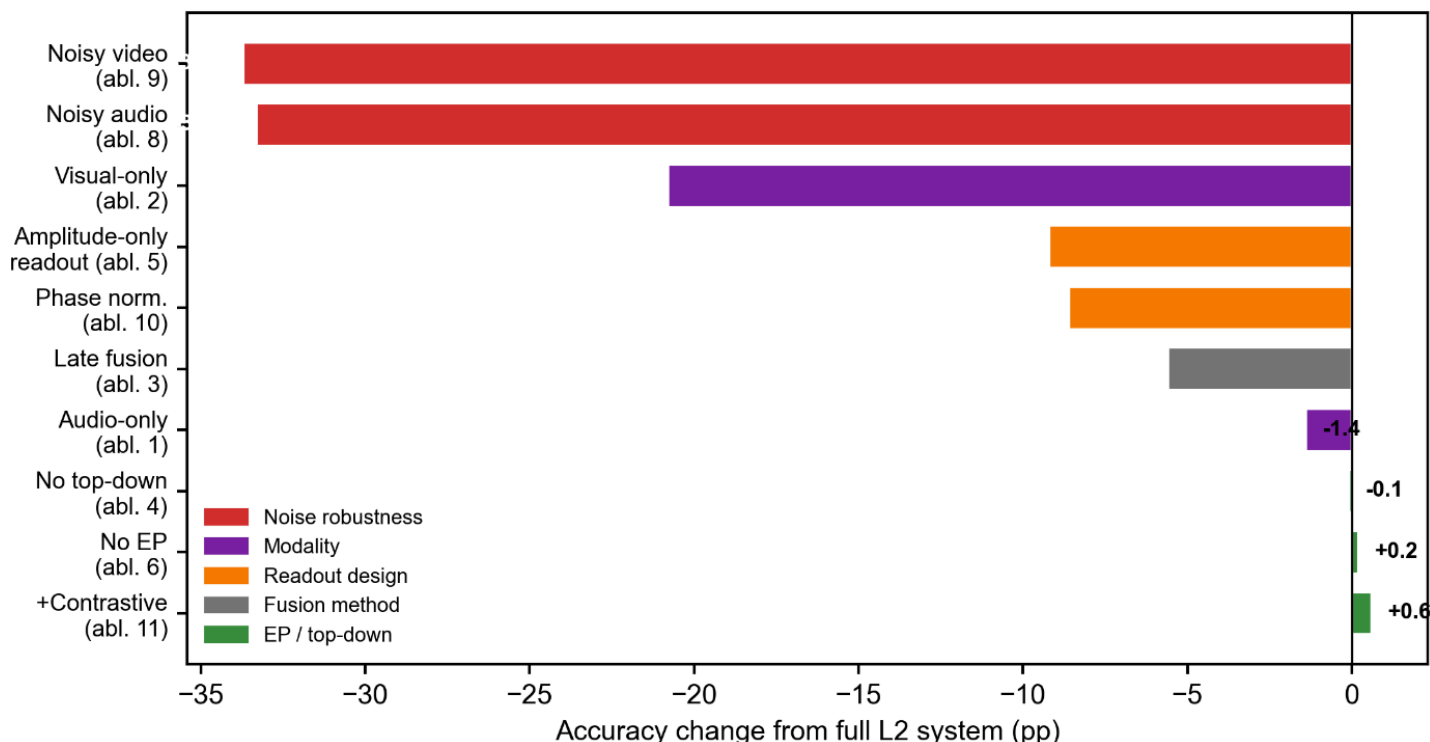


Figure 2

Systematic ablation reveals phase-sensitive readout as the dominant design factor within the L2 system.

Bars show accuracy change relative to the full L2 system control (45.6% validation accuracy) for each ablation condition. Conditions are sorted by impact magnitude. Colors indicate condition category: noise robustness ablations (red, ablations 8–9), modality ablations (purple, ablations 1–2), readout design (orange, ablation 5), fusion method (grey, ablation 3), and EP/top-down ablations (green, ablations 4 and 6). Positive values indicate conditions where removing a component marginally improves accuracy; negative values indicate accuracy loss. The two noise conditions (–33.3 and –33.7 pp for audio and video respectively) and the visual-only condition (–20.8 pp) dominate because they remove an entire stream. Within the remaining conditions, amplitude-only readout (ablation 5, –9.2 pp) is the largest single-component cost, exceeding the analogous unimodal penalty of 7.8 pp reported in the companion paper. Top-down feedback (ablation 4, –0.1 pp) and EP learning of L2 physics (ablation 6, +0.2 pp) contribute within noise.

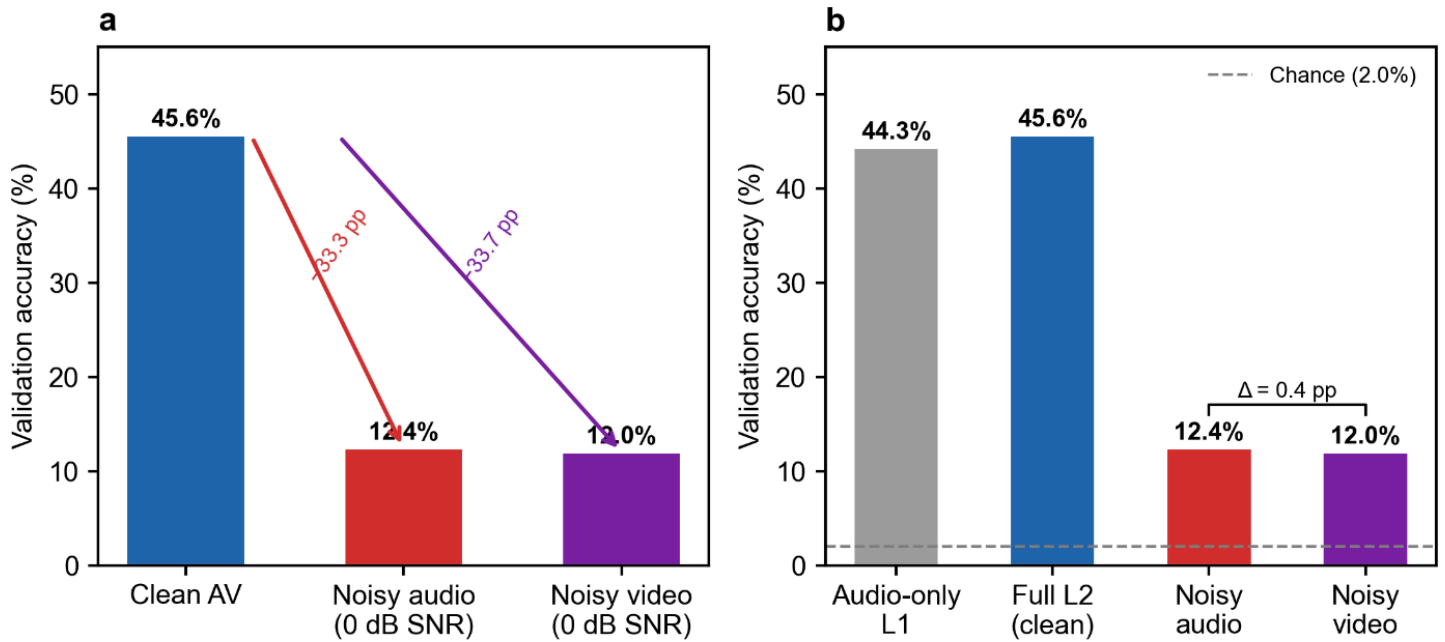


Figure 3

Symmetric noise degradation confirms genuine cross-modal integration.

a, Validation accuracy of the full L2 system under three conditions: clean audiovisual input, audio degraded with Gaussian noise at 0 dB SNR (noisy audio), and video degraded with equivalent noise (noisy video). Arrows indicate accuracy drops relative to the clean condition. **b**, Accuracy comparison across four conditions: the audio-only L1 unimodal baseline (grey), the full L2 system with clean input (blue), noisy audio (red), and noisy video (purple). Brackets annotate the near-identical degradation produced by audio versus video noise (33.3 versus 33.7 percentage points, a difference of 0.4 pp). In a system dominated by one modality, degrading the dominant stream would produce substantially larger accuracy loss than degrading the weaker one. The near-identical drops indicate the L2 binding field distributes reliance across both streams in roughly equal measure. The dashed line indicates chance accuracy (2.0%, 51 classes).

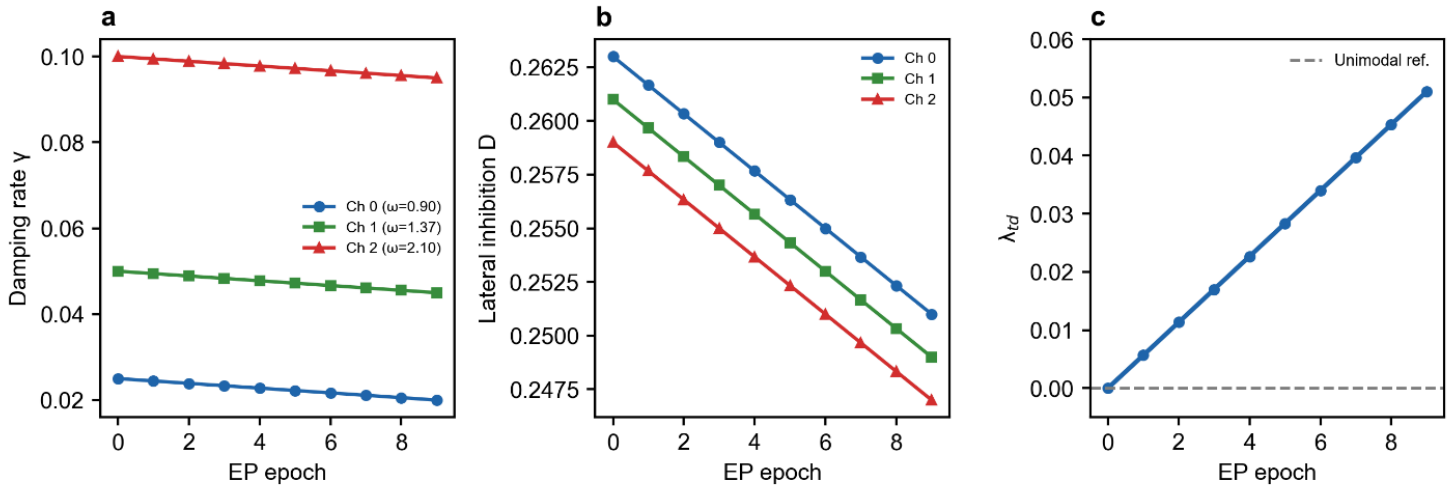


Figure 4

EP drives L2 physics toward lower dissipation, consistent with L1 audio field trajectory.

EP-learned parameter trajectories over 10 training epochs for the L2 binding field. **a**, Damping rate γ for each of the three field channels (channel 0: $\omega = 0.90$ rad/step, blue; channel 1: $\omega = 1.37$ rad/step, green; channel 2: $\omega = 2.10$ rad/step, red). All three channels decreased monotonically (mean $\Delta\gamma = -0.005$ per channel), extending the field's temporal integration window. **b**, Lateral inhibition strength D for each channel. All three channels decreased uniformly ($\Delta D = -0.012$ per channel), tightening spatial coupling in the binding field. **c**, Top-down coupling coefficient λ_{td} (audio and visual channels identical). Growth from 0.0 to 0.051 is shown alongside the dashed reference line at zero indicating the unimodal result. All three parameter types move in the same direction as the L1 audio field reported in the companion paper — toward lower dissipation, reduced lateral inhibition, and extended temporal memory — consistent with a general principle that EP drives LG fields toward low-dissipation regimes when phase structure must be preserved.

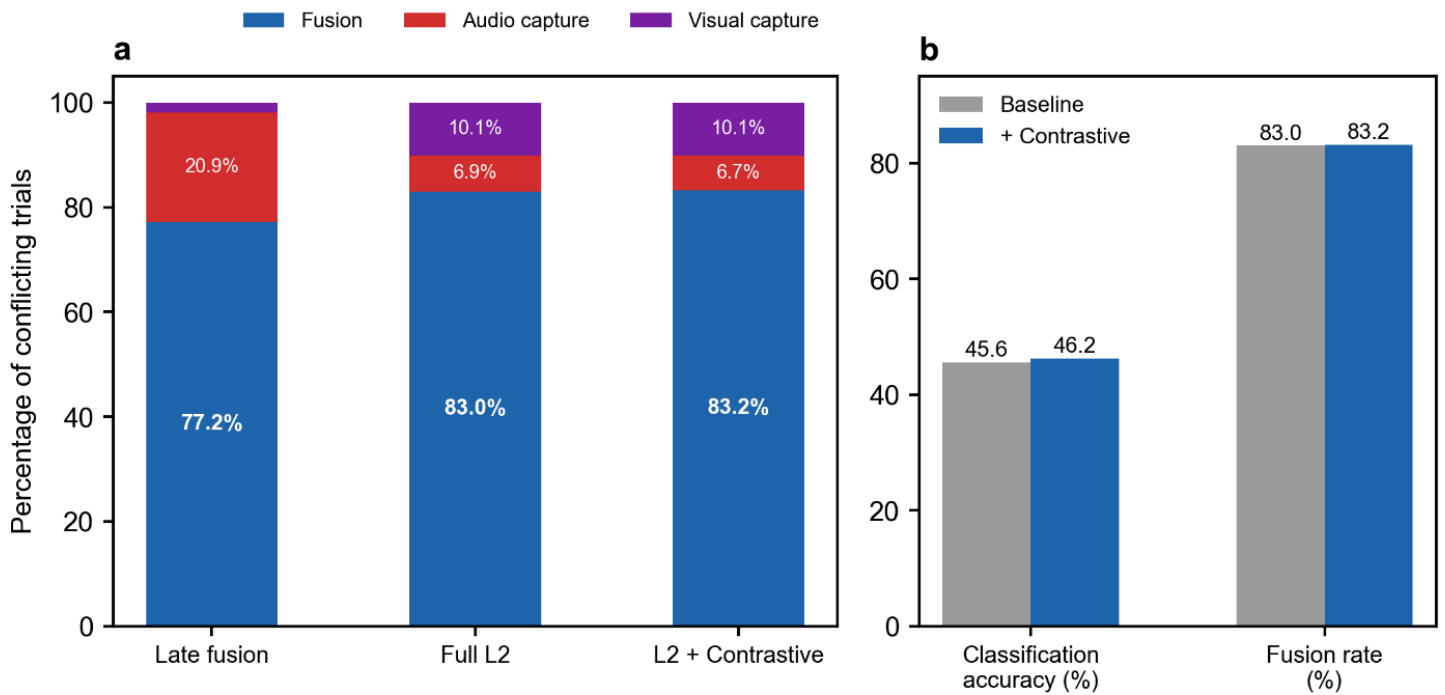


Figure 5

Conflicting audiovisual inputs produce fusion responses reflecting field dynamics rather than readout geometry.

a. Resolution of 4,320 conflicting audiovisual pairs across three conditions: late fusion (feature concatenation without L2 field), the full L2 baseline, and the full L2 system with contrastive readout training (ablation 11). Stacked bars show the percentage of trials resolved as audio capture (red), visual capture (purple), or fusion/neither (blue). The full L2 system produces fusion in 83.0% of conflicting trials, compared to 77.2% for late fusion. Contrastive training shifts the fusion rate by only 0.2 percentage points (83.0% → 83.2%) despite explicitly training the readout to push mismatched representations apart, indicating the 83% fusion rate originates in the L2 field dynamics rather than the readout layer. b. Summary of contrastive training effects. Classification accuracy on matched pairs improves by 0.6 percentage points (grey: baseline, blue: contrastive), while the fusion rate on conflicting pairs is unchanged. The contrastive objective reshaped the feature space without altering the resolution strategy.