

Do vision models perceive illusory motion in static images like humans?

Supplementary Material

Appendix 1 Details on model parameters and architectures.

We evaluated ten motion-estimation models (Table S1):

PWC-Net. A convolutional neural network (CNN) for optical flow estimation through coarse-to-fine refinement. A cost volume is constructed between the first image’s features and the warped second-image features, which a CNN was trained end-to-end to refine the flow estimate [39].

LFN2-K and LFN2-S. *LiteFlowNet2* estimates optical flow by matching local feature neighbors and refining flow through cost-volume correlation and learned regularization, while maintaining a lightweight architecture. The authors provide two trained variants: LFN2-K, fine-tuned for real-world driving scenes (KITTI), and LFN2-S, fine-tuned for rich-textured scenes (Sintel) data [18].

RAFT. *Recurrent All-Pairs Field Transforms* extracts per-pixel deep features from two consecutive frames and constructs a multi-scale correlation volume that contains all pairwise feature similarities. A recurrent module then iteratively updates the flow field by looking up values in this correlation volume and performing optimization. The core idea is to maintain the flow at full resolution throughout iterations (instead of coarse-to-fine), enabling it to recover fine details and handle large motions with high fidelity [44].

CCMR. *Context-guided Coarse-to-fine Motion Reasoning* is a recent optical flow model that integrates attention mechanisms into a multi-scale flow architecture. Global context features are first computed to guide localized motion aggregation at each pyramid level. It has demonstrated strong performance in occluded regions [19].

FlowDiffuser. A novel approach that reframes optical flow as a conditional generation task using diffusion models. It starts from random noise and progressively denoises it into a flow field, conditioned on the two input frames. *FlowDiffuser* can naturally model uncertainty and avoid some training biases of direct regression [27].

FFV1MT. *Feed-Forward V1–MT model* is a human-inspired computational model for optical flow estimation. This feedforward model has two layers that use V1-like spatiotemporal motion-energy filters and MT-like pattern pooling to estimate dense optical flow [37].

DorsalNet. A goal-driven model of the primate dorsal visual stream. This model is a 3D ResNet that was trained to predict an agent’s self-motion parameters from video input. The core idea is that by learning to estimate how an observer moves through the environment, the network develops internal representations similar to neurons in the brain’s dorsal pathway (which is responsible for motion perception) [30].

ME-Attention. *Motion Energy–based Attention model* is a two-stage neural model of human motion perception to combine classical motion energy sensing with modern attention mechanisms. The local signals are fed into a recurrent self-attention network that adaptively integrates motion over space and time. Notably, this model reproduces several neurophysiological and psychophysical observations [40].

Dual. Compared with the ME-Attention model, which uses a single V1–MT motion-energy stream to explain primarily first-order motion perception, the *Dual* model extends this architecture to a dual-pathway system that jointly learns first- and second-order motion. Trained on naturalistic videos with diverse material properties, the Dual model reproduces key psychophysical and neurophysiological findings and achieves dense optical flow and motion segmentation performance comparable to modern computer-vision models [41, 42].

Table S1. Summary of motion estimation models.

Model	Parameters	Input Size	Input Frames
PWC-Net [39]	8.75 M	1024×436	2
LFN2 [18]	6.42 M	1024×436	2
RAFT [44]	5.26 M	520×960	2
CCMR [19]	11.5 M	512×512	2
FlowDiffuser [27]	14.5 M	512×512	2
FFV1MT [37]	N.A.	768×768	5
DorsalNet [30]	57.2 M	768×768	6
ME-Attention [40]	14.7 M	768×768	11
Dual [42]	25.6 M	768×768	15

Appendix 2 Additional Examples of Model Predictions.

Blue–yellow and Red–green variants. The Rotating Snakes illusion induces a robust percept of counterclockwise motion in human observers despite the stimulus being physically static. Figures S1 and S2 show model-predicted optical flow for the blue–yellow and the red–green variants across viewing conditions. Non–bio-inspired models produce minimal or shift-dominated flow, whereas the Dual model exhibits the clearest rotational components under microsaccades, consistent with the trends shown in the main text. For the red–green variant, Dual often produces flow opposite to the expected direction, mirroring the negative correlations reported in the main text.

Illusion vs. control comparisons. Figure S3 directly compares optical flow estimates for illusion and control stimuli across a representative subset of models and shift magnitudes. For non-bio-inspired architectures (RAFT, CCMR, FlowDiffuser), responses to illusion and control stimuli are nearly identical, indicating that the models do not encode the asymmetric luminance gradients critical to the illusion. ME-Attention shows some sensitivity to local structure, but its flow fields remain noisy and lack global coherence. Only Dual consistently differentiates the illusion from its control, particularly under 60–120 px saccades, generating local flow components that align with the expected counterclockwise rotation. Nonetheless, even Dual fails to reproduce the spatial continuity and direction stability observed in human perception, confirming the main-text conclusion that recurrence and dual-channel processing are necessary but not sufficient for full illusory motion replication.

Generalization to stimuli inducing opposite illusory motion perception. To verify that model responses reflect the luminance structure of the stimulus rather than a directional bias, we reversed the luminance order of the micropattern to induce a clockwise percept and re-evaluated all models. Figure S4 shows model-predicted optical flow for stimuli that induce clockwise illusory motion and their corresponding controls. Correlations between model predictions and the expected human illusory percept remain consistent across clockwise and counterclockwise conditions (Figure S4), confirming that the results reported in the main text are not specific to a particular rotation direction.

Appendix 3 Peripheral Viewing.

Peripheral viewing plays a central role in the Rotating Snakes illusion: human observers typically perceive robust illusory rotation when the stimulus lies in the visual periphery or when small eye movements produce transient retinal-image shifts. To emulate these conditions, we tested motion-estimation models on peripheral-viewing simulations in which the stimulus is embedded in a larger uniform field and translated across frames.

Implementation Details. We generated sequences in which a 1386×1386 stimulus was embedded within a 2772×2772 uniform field. Two possible directions were used: bottom-right to top-left, or top-left to bottom-right. In each sequence, the image was shifted by a fixed displacement Δ per update, applied equally in the horizontal and vertical directions. We tested $\Delta \in \{15, 30, 60, 90, 120\}$ pixels, corresponding to the displacement magnitudes utilized in the central simulations described in the main text.

Discussion. Under peripheral-viewing simulations (Fig. S5), none of the models reproduce optical flow patterns consistent with the human percept. Only the bio-inspired architectures correctly identify the disk location and generate coherent rotational flow for the veridical-rotation condition. These results highlight a key distinction between engineering-oriented and neuroscience-inspired motion models. Engineering-oriented optical-flow networks fail under peripheral-viewing conditions because their computations rely on dense feature matching; large textureless regions provide few reliable keypoints, leading these models to produce spurious noise or to default to global smoothness priors. In contrast, models incorporating motion-energy mechanisms exhibit sparse, locally structured responses, reflecting activation of motion-energy units in textured regions and near-zero responses in uniform regions. Although DorsalNet does not include explicit motion-energy units, it contains units with tuning properties resembling biological motion-energy filters.

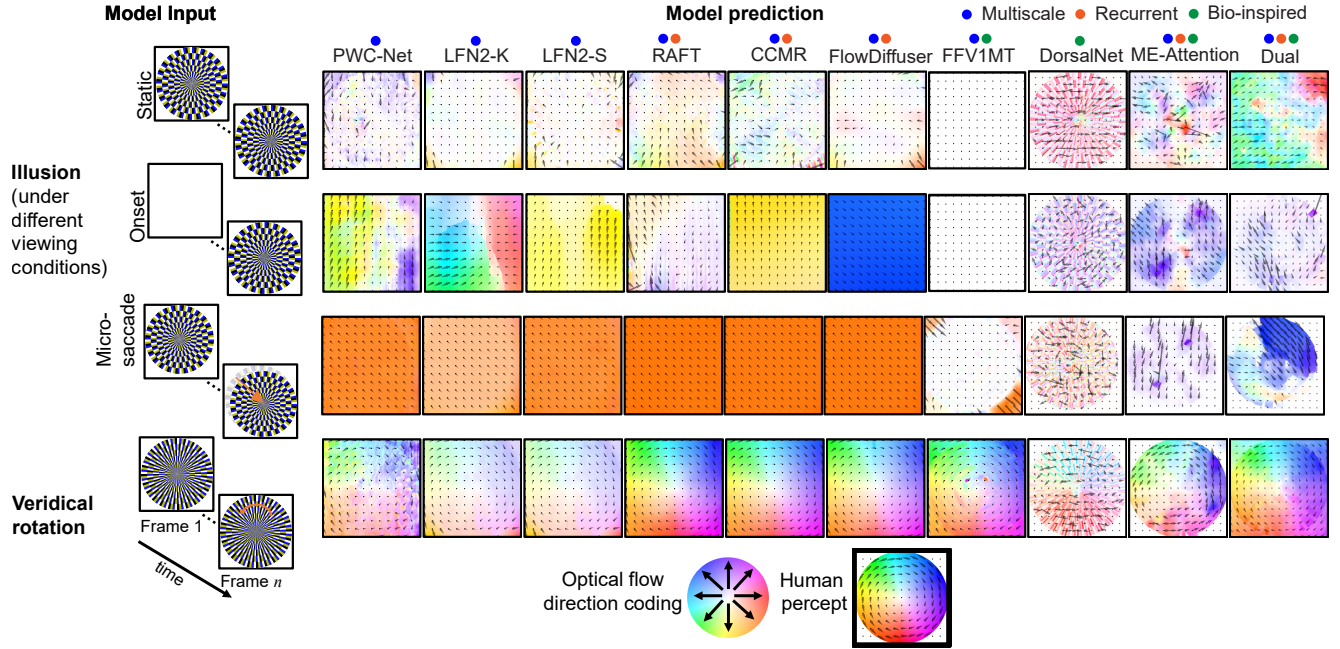


Figure S1. **Visualization of normalized model-predicted optical flow for blue–yellow illusion stimuli.** We evaluated model predictions under three simulated viewing conditions commonly used in psychophysics: (i) *static* presentation, (ii) *onset* presentation, and (iii) *micro-saccade* shifts (small translational displacements that approximate eye movements). For comparison, model predictions for *veridical rotation* of control stimuli are also shown. Colored dots above each column indicate the model architecture type (multiscale, recurrent, or bio-inspired). Optical flow direction is encoded using a circular color wheel (*Optical flow direction coding*, bottom), where hue denotes direction and brightness denotes normalized magnitude. The *human percept* icon (bottom right) represents the expected counterclockwise rotational flow perceived by human observers.

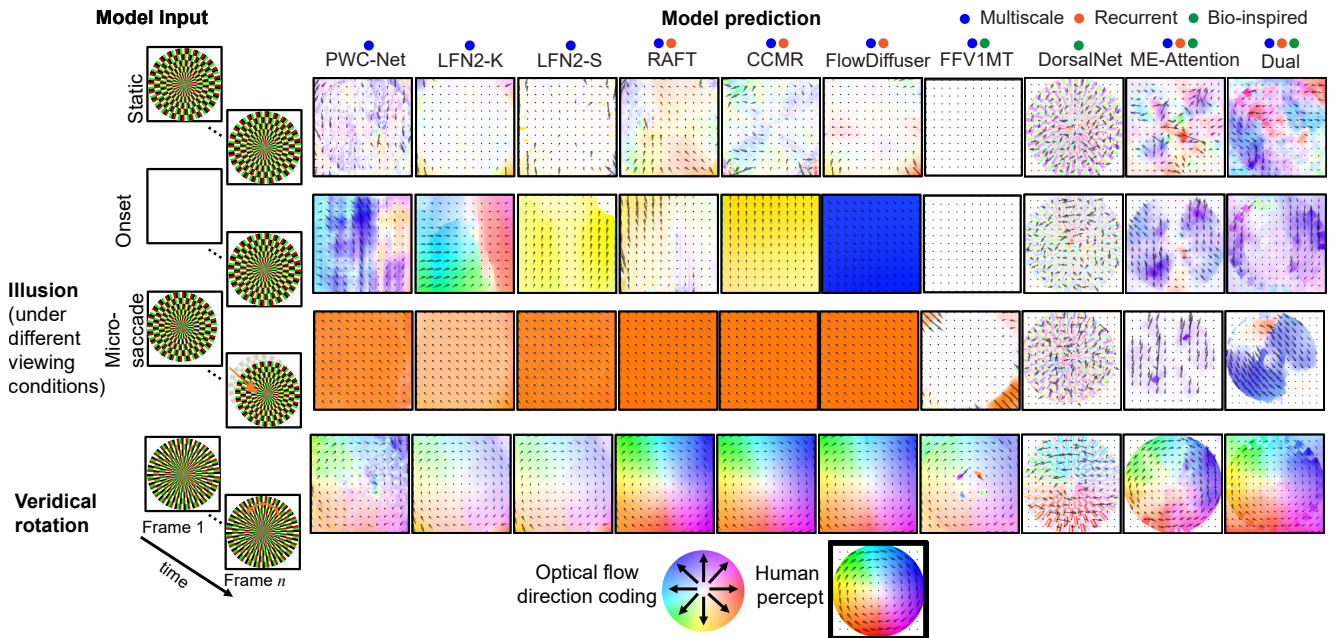


Figure S2. **Visualization of normalized model-predicted optical flow for red–green illusion stimuli.** Figure layout follow the conventions introduced in Figure S1. Here we present the corresponding results for the red–green variant under identical simulated conditions.

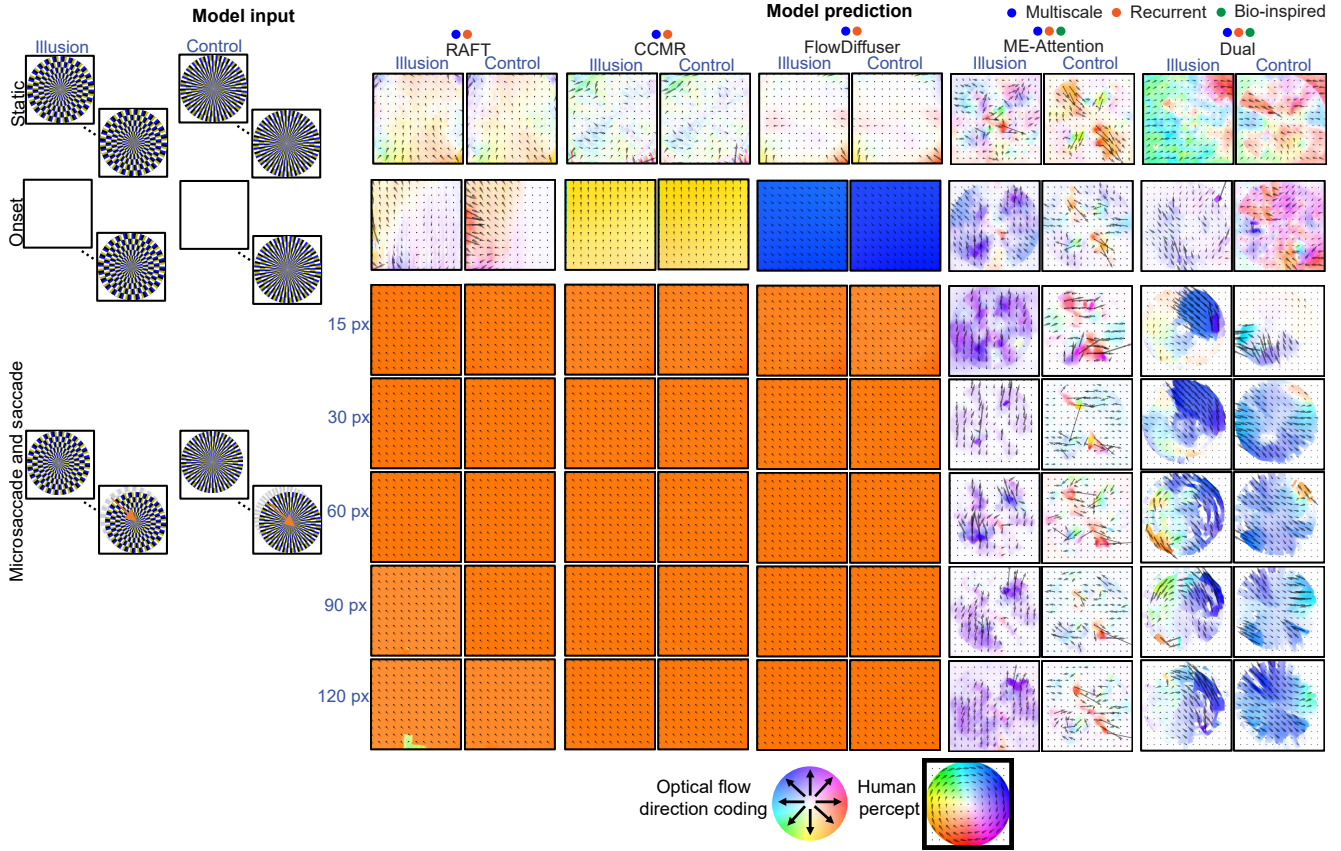


Figure S3. **Visualization of normalized model-predicted optical flow for blue–yellow illusion and control stimuli across select models.** This figure extends the results in Figures 2–3 by directly comparing model outputs for illusion versus control stimuli under matched viewing conditions (static, onset, and microsaccade/saccade shifts of 15–120 px). Results are shown for a subset of representative architectures (RAFT, CCMR, FlowDiffuser, ME-Attention, and Dual). Figure layout, color conventions, and direction-coding icons follow those introduced in Supplementary Figure S1.

Appendix 4 Effect of Timing and Direction of Microsaccade (and Saccade) on Model Perception.

Effects of microsaccade (and saccade) timing. Because non–bio-inspired optical-flow models operate strictly on two-frame inputs, we examined timing effects only for multi-frame bio-inspired architectures. In this analysis, each sequence contained a *single* 30-px microsaccadic displacement, introduced at different possible frames (from frame 2 to the final frame). Consistent with the main-text results, *Dual* is the only model that shows clear positive alignment with human illusory perception (Figure S6). Interestingly, for the red–green variant, *Dual* exhibits pockets of strong positive correlation when the shift occurs near the middle of the sequence, suggesting that the temporal position of the retinal slip can modulate the model’s sensitivity to chromatic asymmetries in the stimulus.

Effects of microsaccade (and saccade) direction. Reversing the microsaccade direction yields qualitative trends consistent with those observed in Fig. 3, with model behavior largely invariant to the shift polarity (Figure S7). As in the main analysis, engineering-oriented models (PWC-Net, LFN2-S, RAFT, CCMR, FlowDiffuser) produce near-uniform or noisy flows dominated by the imposed displacement, and show no evidence of rotation-like organization for either illusion or control stimuli. Only *Dual* shows partially coherent rotation-like flow for the illusion stimulus and results for control stimuli remain near-zero or inconsistent.

We additionally varied the direction of microsaccadic displacements (Figure S8) while holding all other simulation parameters identical to those used in Fig. 3. Eight displacement orientations were tested: 0° , 45° , ..., 315° , with a fixed magnitude of 30 px. Overall, the results are consistent with those reported in the main text: positive correlations for the blue–yellow and negative correlations for

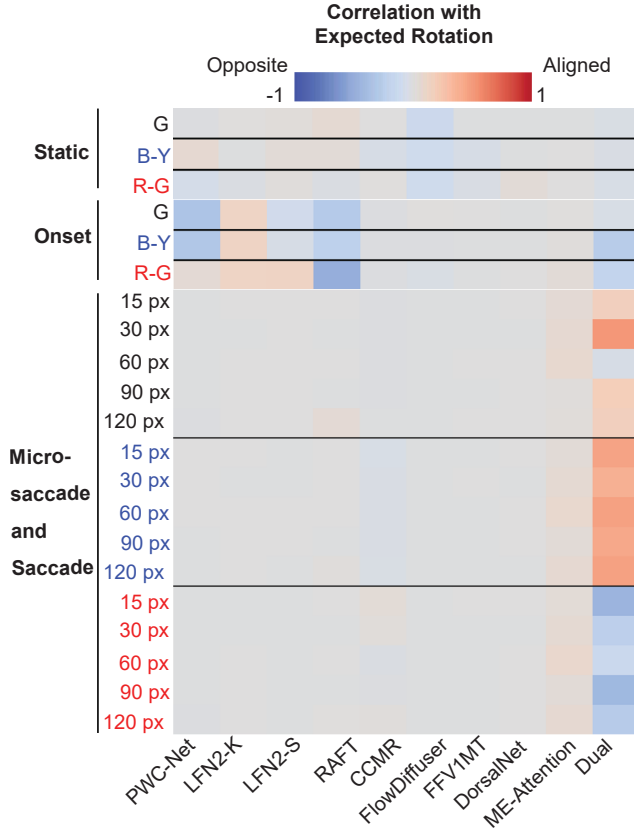


Figure S4. **Correlation between model-predicted and expected human illusory percepts for illusion stimuli inducing clockwise motion perception, and for corresponding control stimuli across models and viewing conditions.**

the red–green variants. Although the model consistently encodes counterclockwise flow for the blue–yellow illusion variant, the *spatial location* at which this rotational pattern emerges depends strongly on the direction of retinal displacement (Figure S8, bottom). These results demonstrate that the Dual model produces illusion effect only in regions where the microsaccade produces sufficiently large local luminance changes along its first- and second-order motion pathways. Because the Rotating Snakes pattern contains directionally asymmetric luminance ramps, different microsaccade directions emphasize different subsets of local units, leading to spatially shifted “islands” of counterclockwise flow. Importantly, control stimuli fail to generate such structured rotational patterns, indicating that the observed responses depend on the luminance asymmetries that drive the illusion rather than on the displacement trajectory alone.

Randomized microsaccade (and saccade) simulations.

We relaxed the experimenter-controlled constraints used in the main text and evaluated the Dual model under more naturalistic retinal-slip statistics (Figure S9). Each sequence contained up to three microsaccades with (i) a random on-

set frame sampled uniformly across the full duration, (ii) a random direction drawn from eight possible orientations $\{0^\circ, 45^\circ, \dots, 315^\circ\}$, and (iii) a random displacement magnitude drawn from $\{15, 30, 60, 90, 120\}$ px. We generated 1000 independent simulations for each color variant (grayscale, blue–yellow, red–green). Model–human alignment was computed as in the main analysis, and statistical significance was assessed using *one-sided Wilcoxon signed-rank* tests evaluating (a) whether illusion correlations were significantly greater than zero, (b) whether control correlations were significantly greater than zero, and (c) whether illusion correlations were significantly greater than control correlations.

Across randomized simulations, the grayscale illusion condition yielded correlations that were significantly less than zero ($p = 3.14 \times 10^{-8}$), and control correlations were non-significant ($p > 0.05$). Consistent with this, illusion and control distributions did not differ ($p > 0.05$). For the blue–yellow variant, illusion correlations were significantly greater than zero ($p = 4.75 \times 10^{-4}$), while control correlations were significantly less than zero ($p = 2.38 \times 10^{-15}$). Illusion correlations were significantly larger than control correlations ($p = 1.21 \times 10^{-15}$), indicating robust human-aligned rotational structure under naturalistic retinal slip.

The red–green variant showed illusion correlations significantly greater than zero ($p = 8.47 \times 10^{-57}$), while control correlations were significantly less than zero ($p = 1.09 \times 10^{-79}$). Illusion correlations were significantly larger than control correlations ($p = 6.97 \times 10^{-120}$). Notably, although controlled-direction microsaccades in Figs. 3 and S8 often produced correlations opposite to the human percept for red–green stimuli, randomized microsaccades yielded predominantly positive alignment.

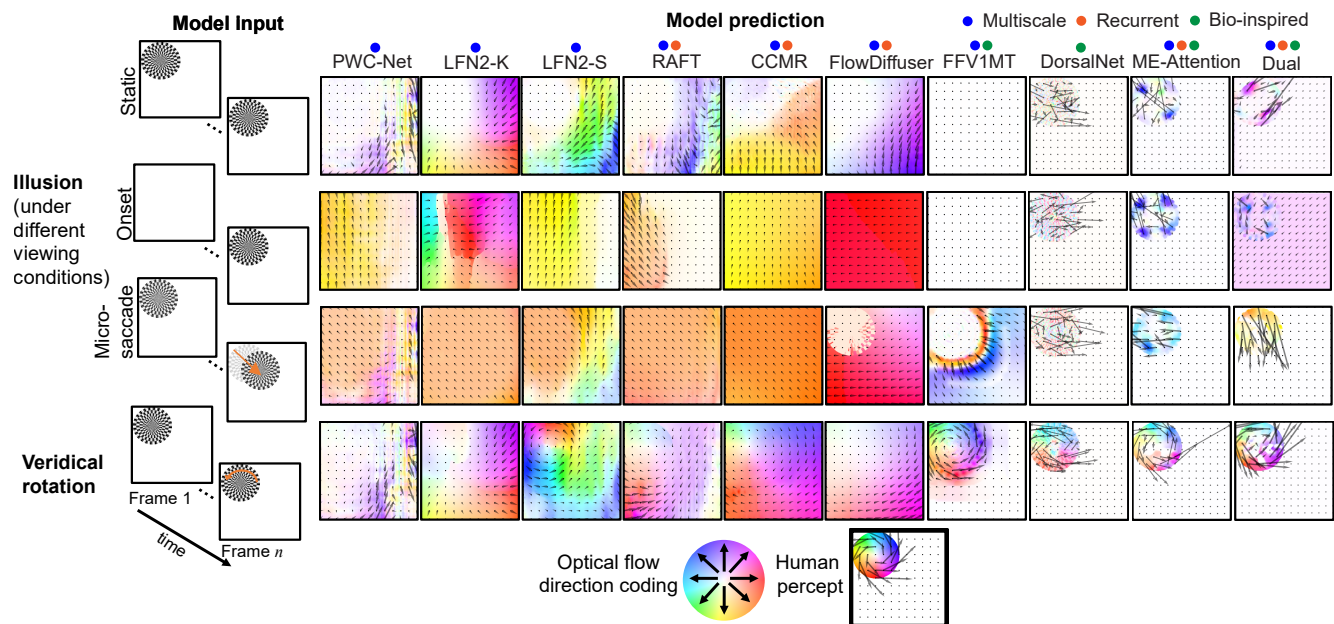


Figure S5. **Visualization of normalized model-predicted optical flow for grayscale stimuli under peripheral viewing condition.** Figure layout, color conventions, and direction-coding icons follow those introduced in Supplementary Figure S1. The “Human percept” icon (bottom right) illustrates the expected counterclockwise rotational flow perceived by human observers. The flow field was defined on the same uniform field and placed at the same spatial location as the stimulus, matching spatial extent and displacement offsets.

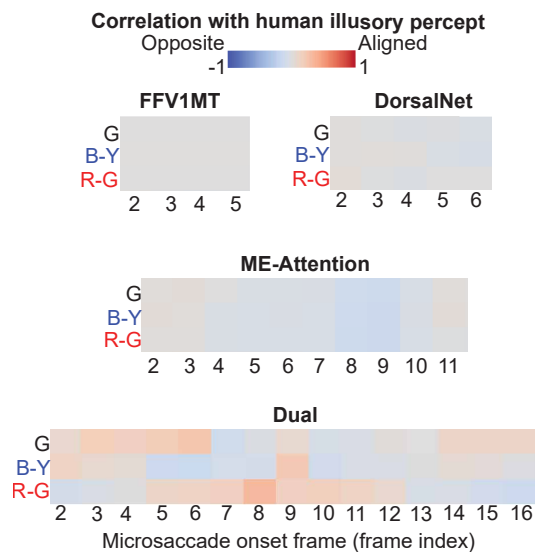


Figure S6. **Correlation across variations in microsaccade timing for bio-inspired models.** For each model, we simulated sequences in which a single microsaccadic displacement (30 px) occurred at different possible time points within the sequence (frame 2 through the final frame). Heatmaps show the correlation between model-predicted optical flow and the expected human illusory percept for grayscale (G), blue–yellow (B–Y), and red–green (R–G) variants.

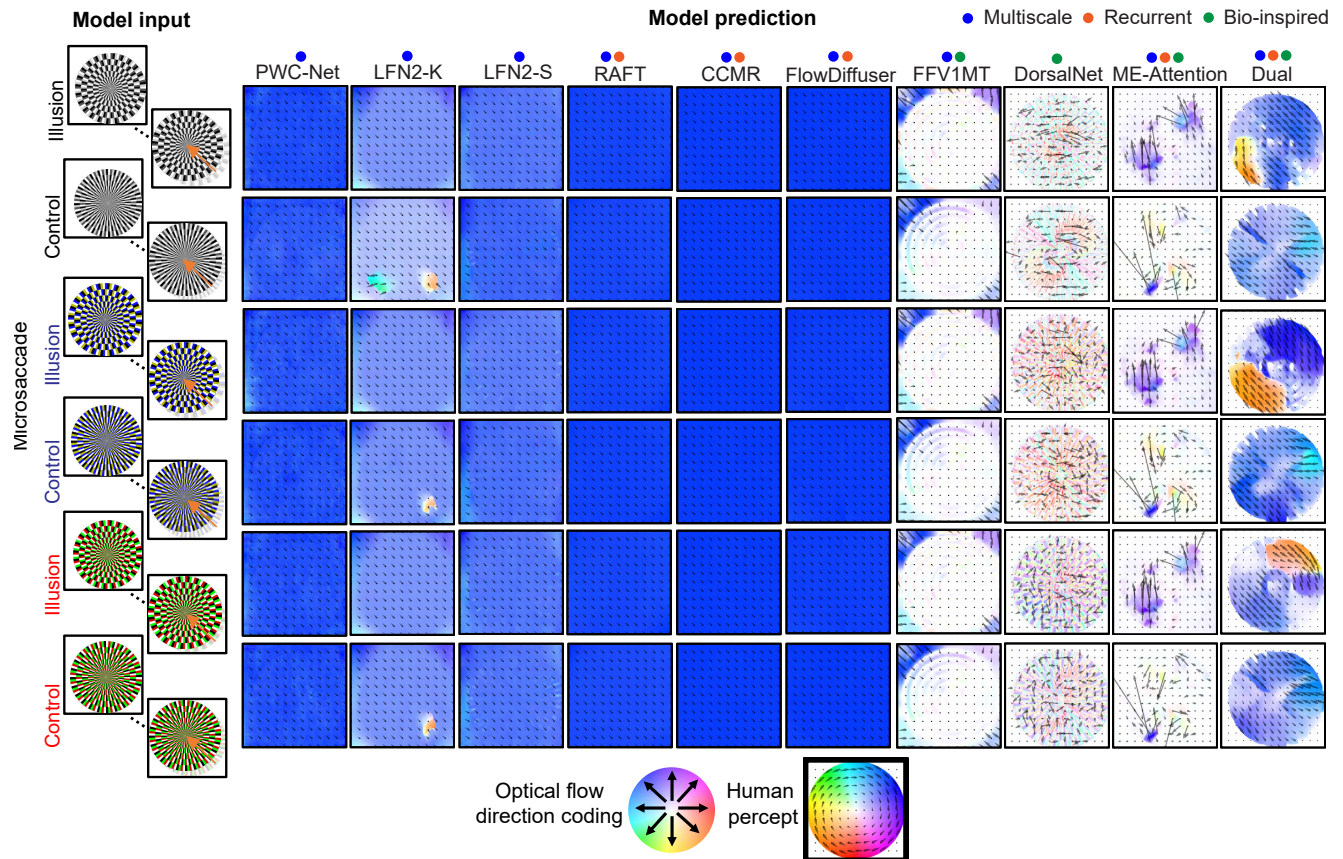


Figure S7. **Effect of reversing microsaccade direction on model-predicted optical flow.** We repeated the central-viewing microsaccade simulation from Fig. 3, but applied the microsaccadic displacement in the *opposite* direction. Model predictions are shown for illusion and control stimuli of three color variants across all models. The bottom icons follow those introduced in Supplementary Figure S1.

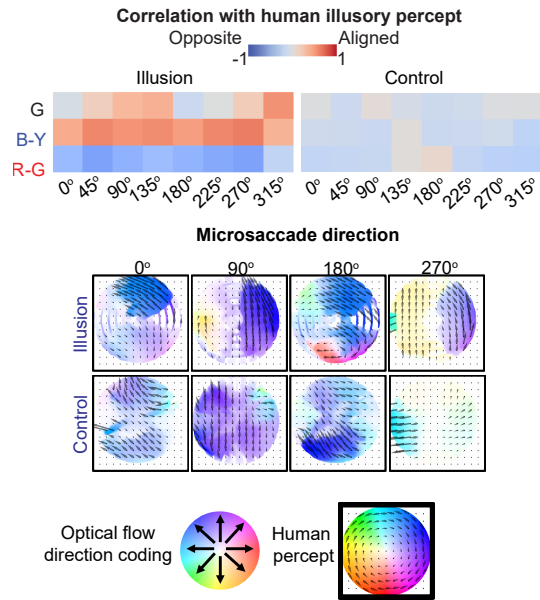


Figure S8. **Effect of microsaccade direction on Dual model predictions.** Correlations are shown for illusion (left) and control (right) stimuli across eight possible microsaccade directions. The direction of the microsaccadic displacement was varied while keeping all other parameters identical to those used in Fig. 3 and Fig. S7. Qualitative flow fields for the blue–yellow variant (bottom) illustrate how Dual’s predicted rotation varies with displacement direction. The bottom icons follow those introduced in Supplementary Figure S1.

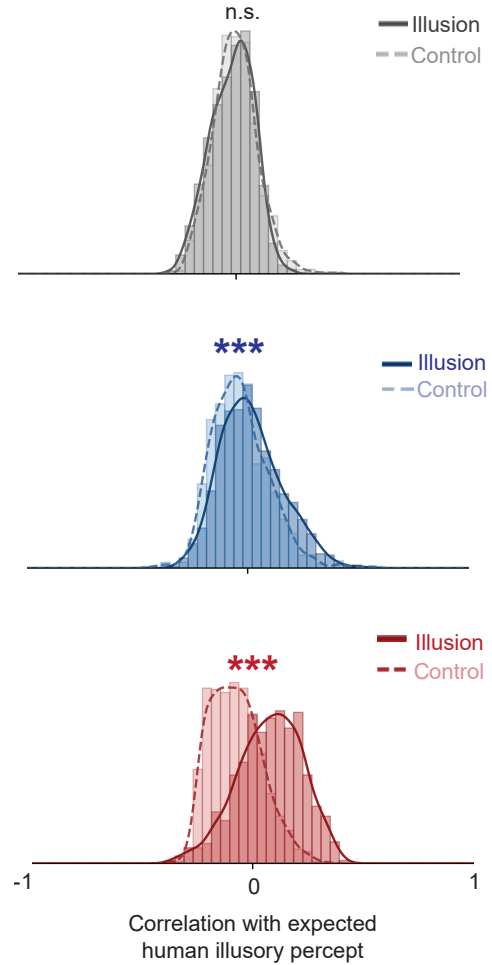


Figure S9. **Distribution of human-alignment scores of Dual model responses under random microsaccades or saccades.** Histograms show the distribution of correlations between predicted optical flow and the expected human illusory percept for grayscale (top), blue–yellow (middle), and red–green (bottom) variants across 1000 simulations. Asterisks denote significant differences between illusion and control distributions (One-sided Wilcoxon signed rank test).

Appendix 5 Other Evaluation Metrics.

For completeness, we report two standard optical flow evaluation metrics: average endpoint error and average angular error [6]. Let $\mathbf{P} \in \mathbb{R}^{n \times 2}$ denote the predicted flow field and $\mathbf{R} \in \mathbb{R}^{n \times 2}$ denote the ground-truth (veridical) flow field at n spatial locations. Each row $\mathbf{P}_{i,:} = (u_i, v_i)$ and $\mathbf{R}_{i,:} = (u_i^*, v_i^*)$ corresponds to the horizontal and vertical components of the flow at spatial location i .

Average Endpoint Error (EPE). The endpoint error measures the Euclidean distance between predicted and ground-truth flow vectors:

$$\text{EPE}_i = \|\mathbf{P}_i - \mathbf{R}_i\|_2, \quad (3)$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm in \mathbb{R}^2 . We report the mean EPE averaged over all spatial locations:

$$\overline{\text{EPE}} = \frac{1}{n} \sum_{i=1}^n \text{EPE}_i.$$

Average Angular Error (AE). Following standard practice, we embed each 2D flow vector (u, v) into 3D space as $(1, u, v)$ to avoid degeneracies when vectors have small magnitude, and we compute the angular deviation between predicted and ground-truth flow as:

$$\text{AE}_i = \arccos \left(\frac{1 + u_i u_i^* + v_i v_i^*}{\sqrt{1 + u_i^2 + v_i^2} \sqrt{1 + u_i^{*2} + v_i^{*2}}} \right). \quad (4)$$

We report the mean angular error averaged over all spatial locations:

$$\overline{\text{AE}} = \frac{1}{n} \sum_{i=1}^n \text{AE}_i.$$

Lower AAE indicates better directional agreement between predicted and veridical flow fields.

Interpretation of alternative metrics. The endpoint error (EPE) and angular error (AE) capture different properties of the flow fields than the correlation metric used in the main text (Figs. S10 and S11). Whereas correlation measures *global directional and magnitude consistency* with the canonical counterclockwise rotational field, EPE and AE evaluate *local vector-wise agreement in magnitude* (EPE) or *directional deviation at each spatial location* (AE), independent of global structure.

These distinctions explain why models such as *DorsalNet* appear to achieve high alignment under EPE and AE despite showing little coherent rotation in Fig. 4. As seen in Figs. S1 and S2, DorsalNet often produces flow fields with several localized regions of strong motion energy, including patches whose directions partially resemble rotational flow. Because EPE and AE are averaged over spatial locations,

these high-magnitude local vectors can reduce the mean error—even when the global pattern does not form a coherent rotation and even when similar rotational components also appear in the control condition.

In contrast, the correlation metric emphasizes whether the overall spatial organization of the flow agrees with the human percept. It penalizes cases where models produce scattered local vectors or noisy fields. Under this more stringent criterion, only Dual captures a globally consistent counterclockwise structure, whereas models like DorsalNet display poorly organized flow that inflates performance under EPE and AE but fails to align perceptually.

Appendix 6 Probing, Control, and Ablation Analysis.

Probing analysis. Fig. S12 shows qualitative examples of optical flow across pathways and recurrent stages. Flow fields decoded from E_1 alone primarily capture local flicker-like motion and lack any coherent tangential structure for the illusion. The addition of the higher-order channel visibly alters the predicted flow fields. The fused representation E_m (Stage I) produces stronger, spatially organized flow for illusion stimuli. This improvement highlights the role of higher-order motion features in generating human-like percepts.

Recurrent integration further amplifies and stabilizes rotational motion. Final-stage outputs (Stage II-6) exhibit globally counterclockwise flow for the illusion condition, closely resembling the veridical rotational field, despite remaining spatially nonuniform. In contrast, control stimuli do not show increases in alignment across iterations, confirming that recurrence selectively enhances features specific to the illusory image rather than simply amplifying input transients.

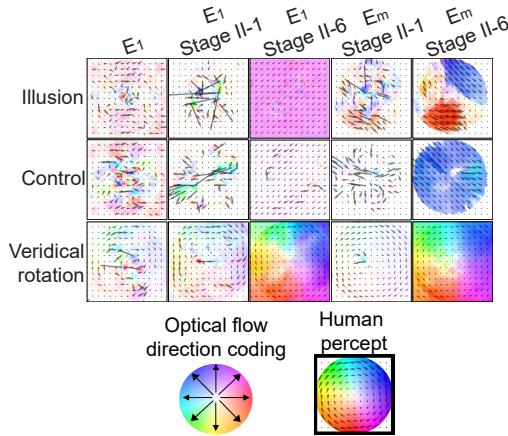


Figure S12. **Visualization of decoded flow fields from representations at different processing stages.** Shown are the first-order pathway alone (E_1), the first and last recurrent iterations applied to E_1 (Stage II-1 and Stage II-6), and the corresponding first and last recurrent iterations applied to E_m .

Control Analysis We retrained RAFT and ME-Attention on the same training data as Dual. Fig. S13 (top panel) shows qualitative optical flow examples under the 30-px microsaccade condition and veridical rotation. RAFT primarily captures the shift direction of the microsaccade, whereas ME-Attention shows some sensitivity to local luminance structure but lacks global coherence, with regions predicting rightward motion. For veridical rotation, both models capture the expected motion, although with reduced global

coherence and weaker speed than models trained on their original training data (see Fig. 3 in the main text).

Ablation analysis Qualitative examples of model-predicted flow after retraining Dual with subsets of the training data and ablating first- and second-order pathways are shown in Fig. S13 (middle and bottom panels, respectively); all variants produced accurate optical flow for veridical rotation. Removing the non-texture object motion training subset leads to noisy local flow inconsistent with global rotation. Removing the diffuse dataset leads to large regions of near-zero flow, with regions of local flow also inconsistent with rotational motion. In contrast, removing the non-diffuse dataset produces predominantly globally coherent but weakly predicted flow, inconsistent with both microsaccade direction and rotational motion. Removing both non-texture and non-diffuse subsets results in flow predictions consistent with the microsaccade direction rather than rotation. Removing the first-order pathway largely eliminates sensitivity to local luminance structure; predicted optical flows are dominated by weak, rightward flow. Removing the second-order pathway produces near-uniform flow with pockets predicting flow in an opposing direction. Predictions for veridical rotation after ablating the first-order pathway align with true rotation, while ablating the second-order pathway introduces regions of weak flow.

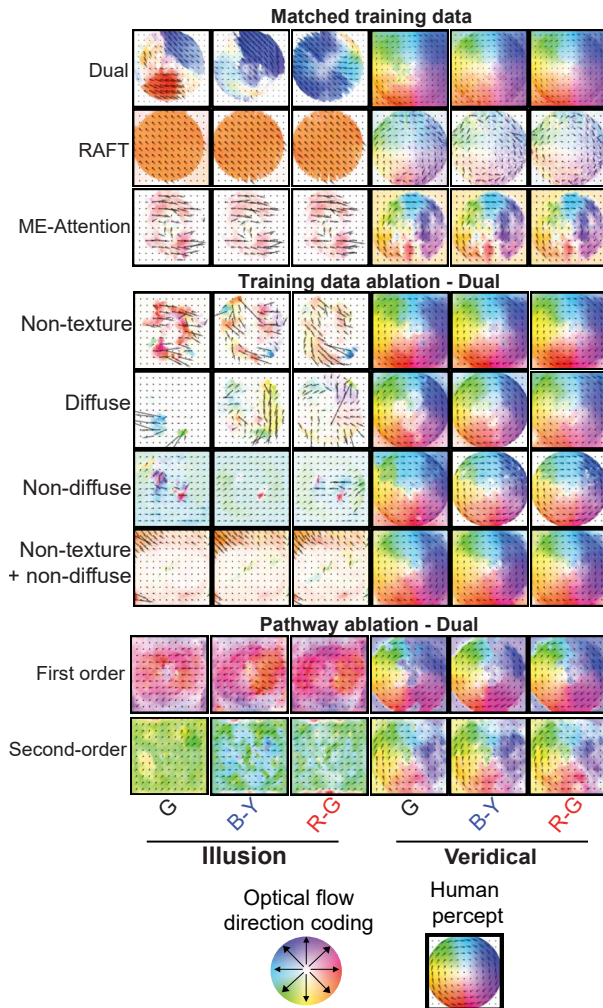


Figure S13. **Predicted flow fields of retrained models evaluated on illusion and veridical-rotation stimuli.** Shown are Dual, RAFT, and ME-Attention trained on the full dataset, as well as Dual trained with data and architectural ablations.

Appendix 7 Other Anomalous-motion Illusion.

We have shown results under random microsaccade simulations in Fig. 7 (see implementation details in Appendix 4). Here, figures S14 and S15 show results under the same simulated viewing conditions used in Fig. 3 (static, onset, and microsaccades).

Peripheral and central drift illusions. All three stimuli (two peripheral drift variants and the central drift illusion) elicit a robust clockwise percept for human observers. Under both the static and onset conditions, none of the evaluated models produced flow patterns resembling this perceptual experience. Consistent with the results in Fig. 7, only the Dual model exhibited any degree of coherent rotational structure under simulated microsaccades, and this effect emerged solely for the peripheral drift illusions. For

the blue–yellow stimulus (middle column), Dual generated a clear clockwise pattern that aligned with the human percept. In contrast, for the grayscale variant (left column), Dual produced a predominantly *counterclockwise* pattern, opposite to the expected direction. The central drift illusion did not induce meaningful rotation in any model.

Ouchi illusion. ME-Attention and Dual exhibit figure–ground segmentation: both models generated distinct motion responses for the central region and its surround even under static and onset presentations. However, the segmentation boundaries were sometimes weak or incomplete—for example, Dual showed degraded segmentation for both the onset and microsaccade conditions. Critically, although ME-Attention and Dual captured coarse region-wise separation, the direction of the predicted flow may not align with the illusory percept experienced by human observers.

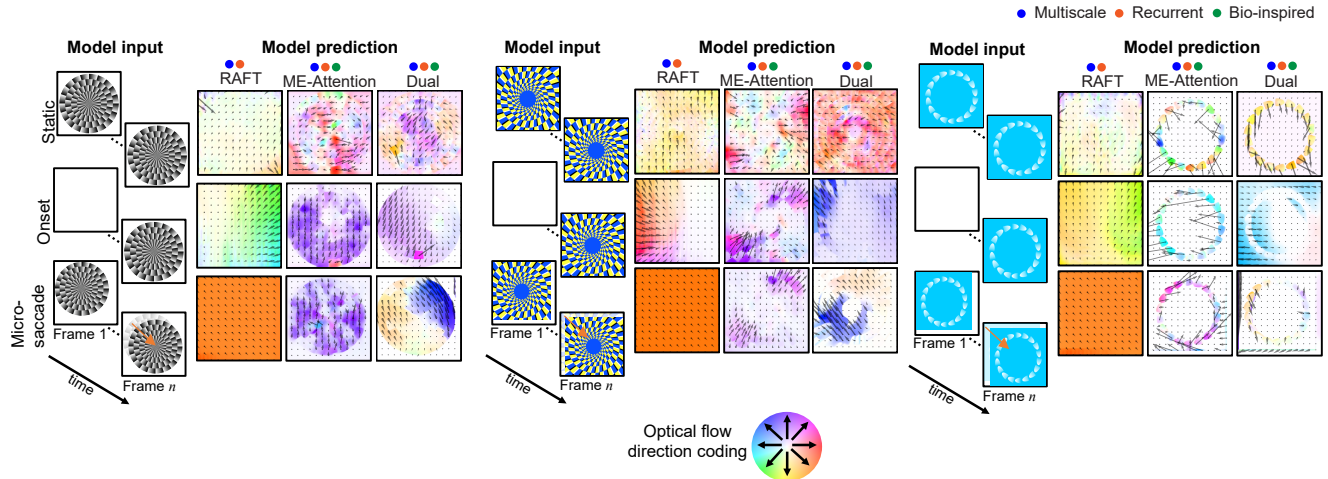


Figure S14. **Visualization of normalized model-predicted optical flow for peripheral drift illusions (PDI) and the central drift illusion (CDI) across selected models.** The layout follows the conventions introduced in Fig. S1. This figure extends the results shown in Figure 6 by presenting the corresponding PDI variants and the CDI under the same subset of simulated conditions as Fig. S1.

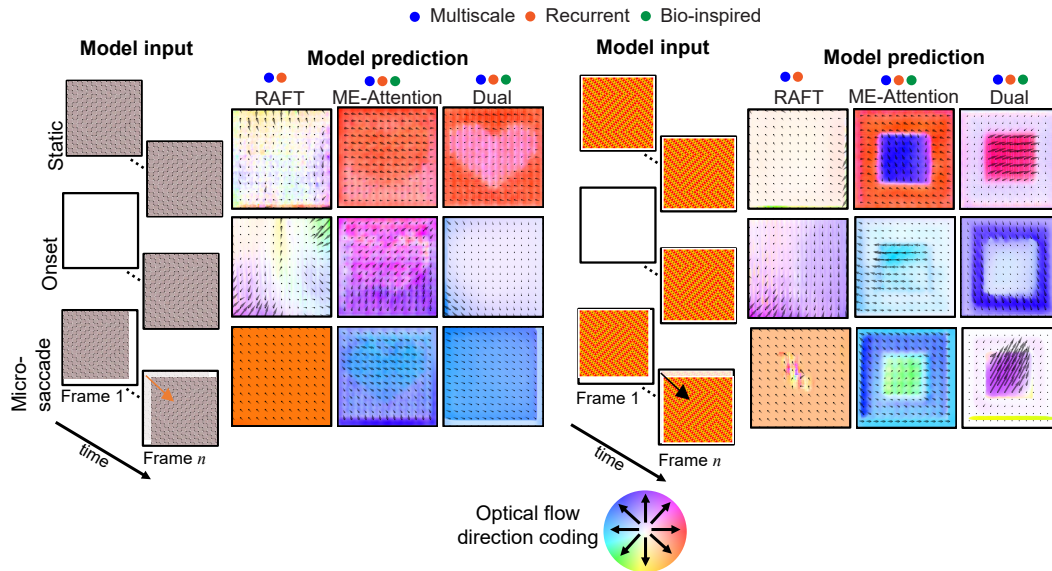


Figure S15. **Visualization of normalized model-predicted optical flow for Ouchi illusions.** The layout follows the conventions introduced in Fig. S1. This figure complements the results in Figure 6 by showing the corresponding Ouchi variants evaluated under the same subset of simulated conditions as Fig. S1.

Appendix 8 Movie Demonstration.

We include video demonstrations illustrating three representative examples of the simulated viewing conditions used as model inputs. All videos are rendered at 5 Hz to aid visualization. Each video includes a fixation cross placed in the bottom-right corner to facilitate stable fixation.

Demonstration of stimulus-onset simulation. Video 1 demonstrates the onset simulation for the grayscale illusion

stimulus.

Demonstrations of microsaccade and saccade simulations. Videos 2 and 3 illustrate the single-shift simulations used for microsaccade and saccade conditions. Video 2 presents a 30 px shift magnitude for the grayscale illusion stimulus, whereas Video 3 shows a 120 px shift magnitude for the blue–yellow illusion stimulus.