

Machine learning modelling for multi-order human visual motion processing

Received: 12 December 2024

Zitang Sun , Yen-Ju Chen , Yung-Hao Yang, Yuan Li & Shin'ya Nishida  

Accepted: 24 May 2025

Published online: 15 July 2025



Visual motion perception is a key function for agents interacting with their environment. Although recent advances in optical flow estimation using deep neural networks have surpassed human-level accuracy, a notable disparity remains. In addition to limitations in luminance-based first-order motion perception, humans can perceive motions in higher-order features—an ability lacking in conventional optical flow models that rely on intensity conservation law. To address this, we propose a dual-pathway model that mimics the cortical V1-MT motion processing pathway. It uses a trainable motion energy sensor bank and a recurrent graph network to process luminance-based motion and incorporates an additional sensing pathway with nonlinear preprocessing using a multilayer 3D CNN block to capture higher-order motion signals. We hypothesize that higher-order mechanisms are critical for estimating robust object motion in natural environments that contain complex optical fluctuations, for example, highlights on glossy surfaces. By training on motion datasets with varying material properties of moving objects, our dual-pathway model naturally developed the capacity to perceive multi-order motion as humans do. The resulting model effectively aligns with biological systems while generalizing both luminance-based and higher-order motion phenomena in natural scenes.

Creating machines that perceive the world as humans do poses a substantial interdisciplinary challenge bridging cognitive science and engineering. From the former perspective, developing human-aligned computational models advances our understanding of brain functions and the mechanisms underlying perception^{1–3}. On the latter side, such models, which accurately simulate human perception in diverse real-world scenarios, would enhance the reliability and utility of human-centred technologies.

Recent advances in machine learning by deep neural networks (DNNs) have led machine vision to surpass humans in performing many vision tasks^{4,5}. In visual motion estimation⁶, state-of-the-art (SOTA) computer vision (CV) models are more accurate than humans at estimating optical flow in natural images⁷; however, they are not yet sufficiently human-aligned, being unable to predict human perception in many aspects. Computer vision models are often unstable under certain experimental conditions^{8,9}. They do not reproduce human visual illusions nor fully capture biases inherent in human perception⁷.

Recent attempts to integrate insights from cognitive science with deep learning techniques^{10–12} demonstrate the DNNs' potential to align with the biological visual motion processing, but they cannot accurately compute the detailed image motion, unlike humans and SOTA CV models.

Here, to contribute both to biological vision science and computer vision, we propose a DNN model showing human-like perceptual responses across broad aspects of motion phenomena, while maintaining high motion estimation capabilities comparable to SOTA CV models.

Our model features a two-stage processing that simulates the cortical system of primates^{13,14}. The first stage mimics the primary visual cortex (V1), featuring neurons with multiscale spatiotemporal filters that extract local motion energy. Unlike past models, the filter tunings are learnable to fit natural optic flow computation. The second stage, which mimics the middle temporal cortex (MT), addresses motion integration and segregation. We introduce the concept of motion

graph modelling dynamic scenes, enabling flexible connections across local motion elements for global motion integration and segregation. As the motion graph implicitly encodes object interconnections in a graph topology, training-free graph cuts¹⁵ can be seamlessly applied for object-level segmentation.

The early version of our model, reported partially in ref. 16, featured a single-channel motion sensing pathway in the first stage and was trained to estimate the ground truth flow across various video datasets. The model successfully replicated a wide range of findings on biological visual motion processing for low-level, luminance-based motion (first-order motion); however, as it is solely based on luminance-based motion sensing, it cannot explain higher-level human motion perception involving spatiotemporal pattern preprocessing, such as second-order motion^{17,18}.

Second-order motion, also termed non-Fourier motion, features high-level spatiotemporal features, including spatial or temporal contrast modulations. Such motion perception is observed across many species, including macaques¹⁹, flies²⁰ and humans^{18,21}, yet it remains undetectable by most CV models⁸. This limitation stems from CV models' reliance on flow estimation algorithms based on the intensity conservation law²², which estimates pixel shifts by matching intensity distributions before and after the movement.

We revised the model's structure and training scheme to encompass both first- and second-order motion perception. As human vision studies suggest separate processing mechanisms for first- and second-order motions^{23–25}, we introduced a secondary sensing pathway with a naive three-dimensional convolutional neural network (3D CNN) preceding the motion energy sensing stage^{23,26}. The 3D CNN is designed to perform nonlinear preprocessing to extract spatiotemporal textures, following the filter-rectify-filter model of second-order motion processing²⁷. Given the computational power of neural networks, the modified model is expected to detect second-order motion after training on an adequate number of artificial, second-order motion stimuli; however, such training is unrealistic in natural environments, where pure second-order motions are rarely observed. The critical scientific question is how and why the biological visual system naturally acquires the ability to perceive second-order motion.

We hypothesized that second-order motion perception aids the estimation of the motion of objects exhibiting different material properties. Natural non-Lambertian optical effects, such as specular reflections and transparent refractions, can alter the light path of an object. This generates complex and dynamic optical turbulence on the surface of the moving object, introducing serious first-order motion noise in the image motion flow. For such non-diffuse materials, detecting first-order motion alone will not provide an accurate estimation of object motion, but the additional use of second-order motion—such as the movements of dynamic luminance noise—would be able to improve the object motion estimation. As a proof of concept that detecting second-order motion correlates with estimating non-Lambertian objects' motion, we created two versions of a motion dataset. One contained purely Lambertian (matte) objects, and the other non-Lambertian objects experienced optical turbulence imparted by non-diffuse materials. We trained different models on both datasets and found that, given an appropriate structure and training environment, the model naturally developed the ability to perceive second-order motion comparable to human capabilities. We also show that our human-aligned visual motion model, with the ability to process both first- and second-order motions, can robustly estimate object motion under noisy natural environments.

The contributions of our study can be summarized as follows:

- To model human visual motion processing by trainable motion energy sensing and a graph network, with the dual-channel design for the detection of both first- and second-order motions.

- To show the model's ability to reproduce past scientific findings related to motion perception while providing high-density optical flow estimation and segmentation comparable with SOTA CV models.
- To demonstrate the conceptual feasibility of a hypothesis that second-order motion perception may have evolved for reliable estimation of motion of non-Lambertian objects despite the presence of optical noise.

Results

In the next section we present the processing pipeline of the dual-channel two-stage motion model. We then demonstrate how the model integrates local motions in various scenarios. Finally, we extend the model's scope to higher-order motions, exploring the relationship between material properties and the ability of second-order motion perception. Demonstrations of our project are available at <https://kucognitiveinformatics.github.io/motion-model-website/>.

The two-stage processing model

Our prototype model features two-stage motion processing that combines classical motion energy sensors in stage I with modern DNNs in stage II. Stage I captures local motion energy, simulating the function of V1, whereas stage II globally integrates local motions, simulating the primary function of the middle temporal cortex. The red route in Fig. 1a is for sensing first-order motion. Specifically, we built 256 trainable motion energy units, each with a quadrature 2D Gabor spatial filter and a quadrature temporal filter. These captured the spatiotemporal motion energies of input videos within a multiscale wavelet space. The key implementation difference from past motion energy models^{13,14} is that we embedded computation in the deep learning framework, with each motion energy neuron's parameters, such as preferred moving speed and direction, being trainable to fit the task. In Fig. 1b we demonstrate the speed–direction distribution and filter receptive field of the trained motion energy neurons. These neurons, activated by stimuli with the preferred spatiotemporal frequency, have their activation patterns decoded into perceptual responses (Fig. 1b(iii)). The activation patterns of stage I resemble mammalian neuron recordings in the V1 cortex with respect to spatiotemporal receptive field and direction tuning (Fig. 1b(ii),(iv)). Moreover, incorporating motion energy sensors allows the model to replicate human-aligned perception of various motion illusions, such as reverse phi and missing fundamental illusions, which are not captured by CV models estimating dense optical flow based on correspondence tracking¹⁶.

Stage I is connected to stage II, which constructs a fully connected graph on local motion energy, treating each spatial location as a node, with all nodes interconnected. We use a self-attention mechanism to define the topological structure of the graph, by which motions are recurrently integrated to generate interpretations of global motion and address aperture problems (Fig. 1a, right). A shared trainable decoder is used to visualize the optical flow fields from stages I and II. The entire model is trained under supervision to estimate pixel-wise object motions in naturalistic datasets^{28–30}.

The first-order motion energy channel can capture first-order motions only. We added an alternative channel to extract information on higher-order motion; this is depicted by the grey route in Fig. 1a. This channel employs trainable multilayer 3D convolutions that extract nonlinear spatiotemporal features before the motion energy computations. This dual-channel design was inspired by earlier vision studies of separate processing designs^{23–25}.

Refer to the 'Model structure' and 'Training strategy' sections in the Methods for more technical details on the model.

Motion graph-based scene integration

This section focuses on how stage II of our model integrates first-order motion signals to solve the aperture problem³¹ by switching off the connection from the higher-order channel in stage I.

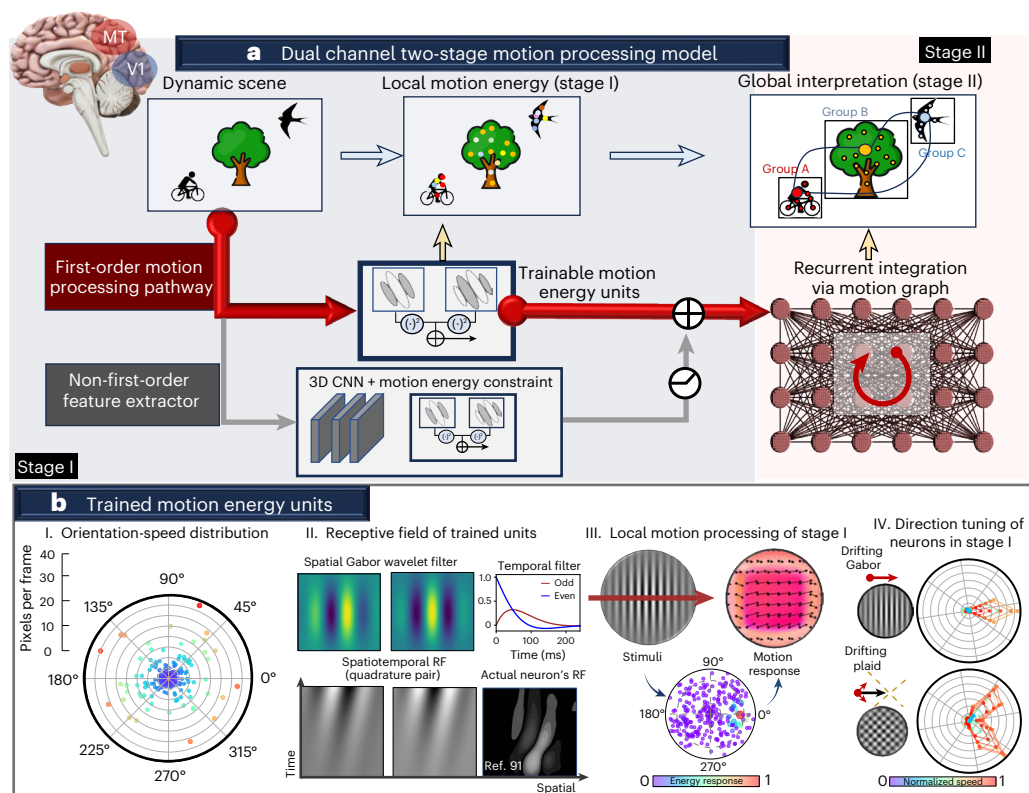


Fig. 1 | Overview of the two-stage motion perception system. a, Stage I mimics the V1 function by detecting local motion, whereas stage II uses a graph network for recurrent motion integration and segregation, mimicking the middle temporal cortex. Stage I employs dual channels to process both first- and higher-order motion. The first-order channel captures Fourier-based motion in motion energy units (red route), and the higher-order channel uses normative 3D CNN layers to extract advanced features (grey route). Natural videos are used to train the entire model for motion flow estimation. **b,** Illustration of motion energy

units in stage I after training. **b(i),** Distribution of preferred moving directions and speeds. **b(ii),** Spatiotemporal receptive field of one motion energy unit, characterized by a pair of Gabor kernel and exponentially decayed sinusoidal kernel. The neuron's receptive field is from ref. 91. **b(iii),** A demo showing how a rightward-moving grating activates specific motion energy units and decodes to a perceptual response. **b(iv),** Direction tuning curves of a specific motion energy unit to grating and plaid stimuli. RF, receptive field.

Figure 2a (left) displays the responses of 256 units to both drifting Gabor and plaid stimuli³². Analysis revealed three distinct groups of units on the basis of their partial correlations with the Gabor and plaid stimuli. Component cells responded to the direction of a Gabor component. Pattern cells responded to the integrated (coherent) direction of plaid motion. Unclassified cells showed no definitive preference for either response, as shown on the right of Fig. 2a. Typically, component cells dominate in V1, whereas pattern cells, equipped with motion integration capabilities, are more common in the middle temporal cortex³². Our model mirrors this biological distribution, as more component cells are in stage I and more pattern and unclassified cells in stage II. Figure 2b shows a global motion of drifting Gabors, where each local patch exhibits a different local direction and speed but is collectively consistent with unified 2D motion downward. Humans perceive coherent downward motion by integrating local motions across space and orientation³³. In agreement with human perception, stage I of our model computes local motion whereas stage II responds to global motion.

Figure 2c illustrates how the model adapts to spatial patterns when integrating motions. When a diamond moves along a circular path (scenario A), where stage I would detect local orthogonal movements of the line segments, stage II integrates the local motions into a coherent global motion (see the left side of Fig. 2c). In scenario B, despite the corners of the diamond being occluded by stationary rounded squares, the model integrates the local motions of the line segments into a single coherent motion. The heat map of the stage II connections shows that the line segments remain linked, as if the model properly considers

the spatial relationships between occluders and edge segments. This cannot be simply attributed to a wide integration window from the motion graph because, in scenario C, where the occluders are invisible, the connections between the line segments are lost in stage II, and the model generates incoherent motion. These model behaviours across scenarios A–C align well with human psychophysical data³⁴, as shown by the similarity in the motion coherence index between the model and humans (see bar plot at the bottom-left of Fig. 2c).

Stage II is essential when processing complex natural scenes (Fig. 3a). Real scenes often exhibit chaotic local motion energies, compounded by challenges such as occlusions and non-textured regions. Addressing these complexities requires long-range and flexible spatial interactions, which are effectively handled by the graph-based, recurrent integration process of stage II. During the iterative process, the model represents local regions as nodes of a graph. The connection weights between locations are captured by the adjacency matrix $A \in \mathbb{R}^{HW \times HW}$. This matrix is normalized to within the range (0,1), where higher values indicate stronger connections. An affinity heat map can be expanded from a specific row of the adjacency matrix (Fig. 3a), indicating how stage II distinguishes objects from the background and adaptively establishes connections across occlusions. We hypothesized that some of the information required for object-level segmentation was inherently encoded in the topology of the motion-based graph. We used a training-free visualization method to test this. Specifically, graph bipartitioning based on the eigenvector corresponding to the second smallest eigenvalue of the graph Laplacian¹⁵ enabled instance segmentation based on motion coherence (right side of Fig. 3a).

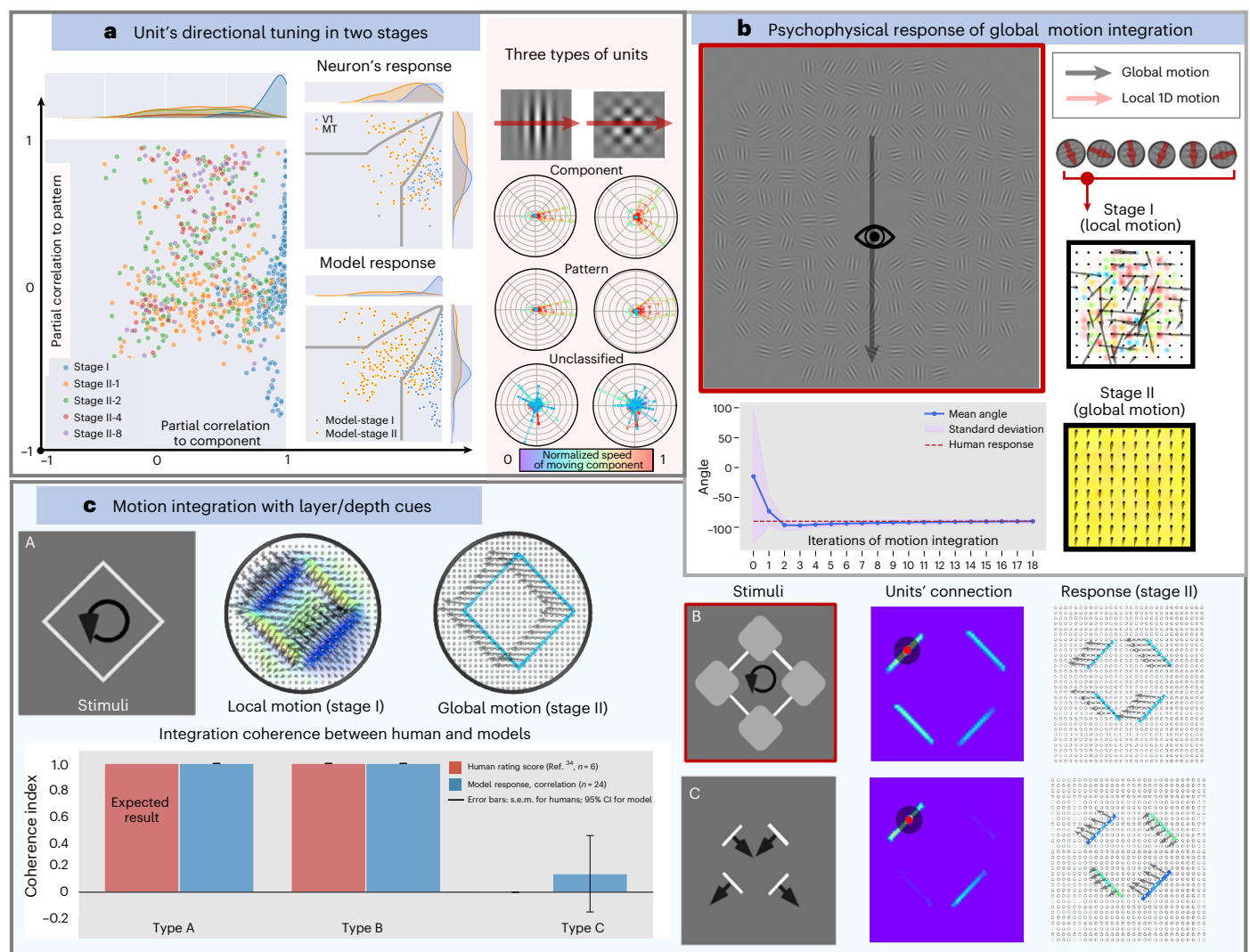


Fig. 2 | Recurrent motion integration. **a**, The tuning properties of model units for 1D gratings and 2D plaids. The partial correlations to component and pattern tuning types are shown. Overall, the model units exhibit a trend similar to mammalian neural recordings: Stage I predominantly consists of component-selective neurons, whereas stage II is dominated by pattern-selective neurons, corresponding to the V1-MT processing hierarchy. The animal data are from ref. ³². **b**, Response to global motion of Gabor patches³³. Local patches contain various motion directions and speeds, which are captured by stage I. Stage II then performs motion integration, linking local motion signals to resolve the aperture problem and infer global (downward) motion. The model response

aligns with the human perception of adaptive pooling. **c**, Motion integration is sensitive to higher-order pattern cues. We used the three scenarios A, B and C detailed in ref. ³⁴. The extents of integration were quantified by correlating the directions of motion between adjacent segments across a single circular translation cycle. Compared with scenario C, scenario B—characterized by structural constraints and depth cues—led to an increased integration index in the model, similar to human perception³⁴. In the middle column of the right panels (for unit connections), we visualize the attention heat map derived from the motion graph, showing the connectivity of the unit (marked by a circle) with other units in stage II.

The results indicated that the model integrated motion representations and object-level recognitions via graph structure, grouping objects even across occlusions.

Our motion-graph-based integration mechanism can unify motion perception and object segmentation in a single framework. Through a recurrent process, local motion signals become accurately combined in a graph space, yielding clear object-level representation in a coarse-to-fine manner. Refer to the 'Stage II (global Motion integration and segregation)' section for the implementation details. This may be related to motion-shape interactions in the biological visual system³⁵.

We further tested the model using the Sintel slow benchmark²⁸, for which psychophysically measured human-perceived flows are available⁷. We compared our model to various CV optical flow estimation methods, including traditional algorithms such as Farneback³⁶; biologically inspired models^{11,37}; and SOTA CV models such as multiscale

inference methods^{6,38}, spatial recurrent models³⁹, graph reasoning approaches⁴⁰ and vision transformers⁴¹. As detailed in Table 1, we computed the Pearson correlation coefficients and vector endpoint errors (EPEs) to assess the relationships between model predictions, human responses and ground truth. We also calculated partial correlations between human and model responses while controlling for the influence of ground truth, and the response consistency index (RCI)⁷. These two are global and local measures to evaluate how much the model prediction accurately replicates human perceptual errors from the physical ground truth (refer to the 'Human and model comparison' section for further details).

Although our framework was not explicitly optimized for precise flow estimation, its performance remains competitive with SOTA CV models. Notably, our model shows the highest partial correlation with human response and RCI. Figure 3b demonstrates a strong

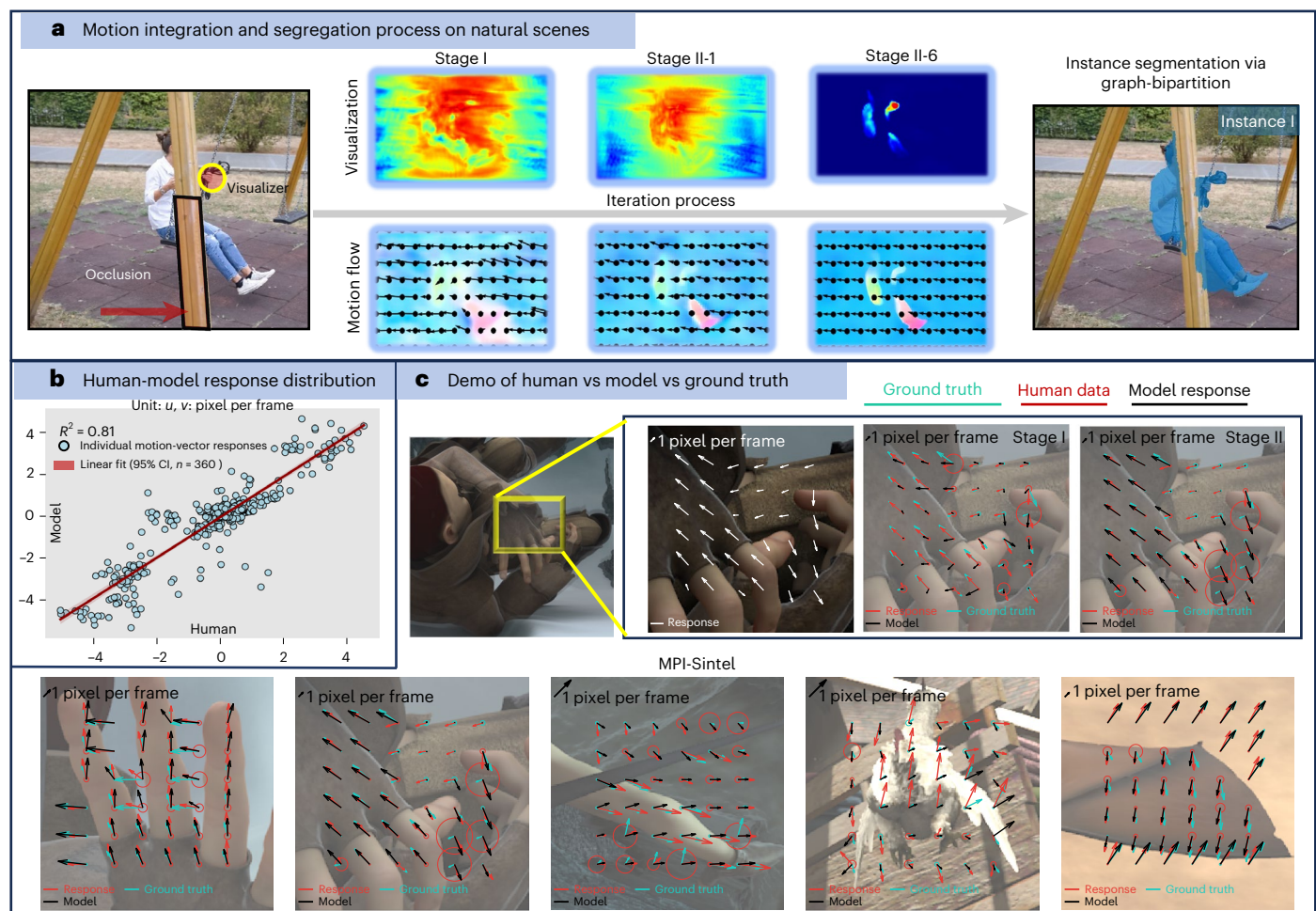


Fig. 3 | Recurrent motion integration in naturalistic scenarios. **a**, Stage II- N refers to the outcomes from the N th iteration in stage II, with visualizations illustrating neuron connectivity via heat maps. The neural connectivity is represented as a graph structure, and through graph bipartitioning¹⁵ the model can further achieve instance segmentation without any additional training. **b**, A scatter plot comparing human and model responses to the Sintel dataset in terms of u, v (pixels per frame). A red regression line, fitted to the data, indicates a strong linear relationship between human and model responses. The shadows

around the red line represent the 95% confidence interval (CI) of the linear regression line. **c**, Qualitative comparison of ground truth, human and model responses on the MPI-Sintel²⁸. The larger the red circle at each location, whose size indicates the magnitude of positive RCI, the stronger the alignment with human responses over ground truth. At many points, our model demonstrates a better alignment with human perception than with ground truth. See Table 1 for detailed quantitative results. The model only uses the first-order channel in stage I, vs. versus.

correlation between the model prediction and human data in (u, v) vector component distribution. Figure 3c qualitatively suggests that motion integration in stage II introduces perceptual biases that align with human errors.

In addition to the Sintel benchmark, we tested our model on the KITTI 2025 dataset³⁰, which consists of real-world driving scenes, and found consistent results (Extended Data Table 1). See also Extended Data Table 4 for the results obtained when the dual channels are used on these benchmarks.

Material properties and second-order motion perception

In this section we will consider a full dual-channel model. Despite including a second channel that extracted higher-order features, our model could not identify second-order motion when trained only on existing motion datasets. This limitation reflects broader challenges in CV, as other DNN-based models also fail to capture second-order motion perception⁸.

To test our hypothesis that the biological system evolved to perceive second-order motion for estimating object movement amidst optical noise from non-diffuse materials, we constructed datasets that controlled the properties of object materials. One dataset contained

diffuse (matte) reflections and the other non-diffuse properties, including glossy, transparent and metallic surfaces (Fig. 4a). The model was trained with a focus on higher-order motion extractors to estimate the ground truth of object motion while ignoring optical interferences caused by non-diffuse reflections.

To quantify second-order motion perception, we developed a benchmark using natural images with various second-order modulations. As shown in Fig. 4b, the benchmark included classical drift-balanced motion (temporal contrast modulation)¹⁷; local low contrast (spatial modulation); and natural phenomena such as water waves and swirling flow fields (spatiotemporal modulation). The last movements are not pure second-order motion but are almost indiscernible in Fourier space, given the chaotic optical disturbances caused by reflection and refraction. Our psychophysical experiment revealed a strong correlation between the physical ground truth and the human response in detecting second-order motion ($r_{\text{mean}} = 0.983$, s.d. = 0.005) (Fig. 4c). By contrast, a representative CV model, RAFT, was associated with a much lower correlation ($r = 0.102$). We trained our model on the diffuse and non-diffuse datasets and compared the correlations with human responses. The results of Fig. 5c indicate that both the dataset material properties and the model architecture greatly influence

Table 1 | Model versus human versus ground truth on Sintel benchmark

Method	ρ_{uv}	ρ_{dir}	ρ_{spd}	RCI	vs Human				vs Ground truth			
					r_{uv}	r_{spd}	r_{dir}	EPE	r_{uv}	r_{spd}	r_{dir}	EPE
Farneback ³⁶	0.27	0.23	0.11	0.039	0.41	0.91	0.34	2.02	0.34	0.33	0.92	1.96
FlowNet2.0 ⁶	0.39	0.26	0.34	0.034	0.92	0.90	0.96	0.94	0.95	0.94	0.98	0.47
RAFT ³⁹	0.20	0.22	0.14	0.026	0.92	0.90	0.96	0.93	0.98	0.99	0.99	0.25
RAFT-val	0.43	0.17	0.42	0.049	0.92	0.89	0.96	1.01	0.92	0.89	0.98	0.69
AGFlow ⁴⁰	0.30	0.16	0.20	0.016	0.93	0.90	0.96	0.92	0.98	0.98	0.98	0.27
GMFlow ⁴¹	0.34	0.32	0.17	0.028	0.91	0.84	0.96	1.03	0.93	0.90	0.97	0.73
FlowFormer ³¹	0.36	0.14	0.32	0.030	0.93	0.91	0.95	0.90	0.98	0.97	0.98	0.42
FFV1MT ³⁷	0.31	0.16	0.31	0.043	0.83	0.64	0.92	1.48	0.59	0.84	0.94	1.29
DorsalNet ¹¹	0.17	0.19	−0.10	0.029	0.20	−0.08	0.86	2.35	0.20	−0.04	0.86	2.33
Ours-first	0.50	0.38	0.37	0.067	0.91	0.88	0.95	0.93	0.90	0.90	0.96	0.88

Text marked bold indicates the best performance in that column. ρ , Partial correlation between human and model controlling for ground truth; r , Pearson correlation coefficient; EPE, vector endpoint error; uv, dir and spd represent motion components in Cartesian space, direction and speed, respectively; RCI denotes model–human similarity (larger = more human-aligned). Ours-first uses only the first-order channel.

the perception of second-order motion. Even when trained with our non-diffused data, the tested CV models still show a limited capability to recognize second-order motions. By contrast, our dual-channel model, trained with non-diffuse data, substantially improved recognition of second-order motion. The average correlation reaches 0.902 (right side of Fig. 5c).

Figure 5b shows the directional tuning capacities of the first- and higher-order motion channels. For various directions of first- and second-order drifting gratings, directional tuning was estimated using the modified circular variance⁴². The first-order channel responded primarily to first-order motions, whereas the higher-order channel was more sensitive to second-order motion. The sensitivity of the higher-order channel to the second-order motion was further enhanced through training on non-diffuse materials (compare red and blue dots in Fig. 5b).

We also compared the Pearson correlations between our final model responses and motion ground truth across SOTA optical flow models, including RAFT, GMFlow and multi-frame-based VideoFlow⁴³. As shown in Fig. 5a, our model exhibited the highest correlation and stability, closely matching human performance. Extended Data Tables 2 and 3 provide more detailed quantitative data on second-order motion comparison.

Notably, unlike our dual-channel model, SOTA CV models cannot achieve a good ability to detect second-order motions even after training with non-diffuse materials. This limitation probably stems from structural design. Computer vision models are primarily designed to track the absolute pixel correspondences between frames, and thus rely on pixel intensity⁴⁴. As second-order motions such as drift-balanced motion lack explicit pixel correspondences across frames, such models often become unstable and generate noisy responses.

The interplay between the first- and higher-order channels

Extended Data Fig. 1a presents qualitative data illustrating the difference between the first- and higher-order channels, demonstrating their function when processing natural scenes with noisy optical environments (first row). Higher-order processing affords more stable results when interpreting global flow motion (left). Such processing effectively tracks the movement of a plastic box with fluctuating water inside, even outperforming certain SOTA CV models⁴⁵ when handling such extremely noisy—but natural—scenes. The second and third rows show the segmentation results for both natural scenes and pure drift-balanced motion¹⁷. In terms of segmentation, the higher-order channel usually helps the model to identify objects in motion. The segmentation results are finer than those of the first-order channel

alone. We validated these results on the DAVIS 2016 video segmentation benchmark⁴⁶, which includes 3,505 image samples. The dual-channel approach achieved a mean intersection over union (IoU) score of 0.60, outperforming the single-channel method, which scored 0.56. In the last row of Extended Data Fig. 1a, we show that our framework can group objects, even when they are spatially invisible, as seen in the pure drift-balanced motion test. The higher-order channel affords a distinct advantage under such conditions, effectively identifying object instances within noise. Such second-order motion patterns are near-undetectable by current CV segmentation models, including SOTA video segmentation models^{47,48}. Note that our segmentation results were obtained using a naive graph bipartition¹⁵ without additional training. Refer to the ‘Stage II (global motion integration and segregation)’ section for implementations of the motion graph.

Discussion

We establish a human-aligned optic flow estimation model capable of processing both first- and higher-order motions. The model replicates the characteristics of human visual motion in various scenarios ranging from typical stimuli to more complex natural scenes.

Recent studies have also leveraged DNNs to infer the neural and perceptual mechanisms underlying visual motion. For example, Rideaux et al.^{10,49}, and Nakamura and Gomi¹² used multilayer feed-forward networks, whereas Storrs and colleagues⁵⁰ used a predictive coding network (PredNet) to model human visual motion processing. DorsalNet¹¹ employed a 3D ResNet model to predict self-motion parameters. Despite their contributions, these models cannot estimate the dense optical flows consistent with the physical or perceptual ground truths, nor do they account for higher-order motion processing.

Modelling visual motion processing

We modelled human visual motion processing, including the V1-MT architecture, via motion energy sensing and graph-based integration. After end-to-end training, our model generalized both simple laboratory stimuli and complex natural scenes well. The model naturally captures various characteristics of neurons in the motion pathway, including the change in spatiotemporal tuning from the V1 to the middle temporal cortex areas. Motion integration successfully explains the physiological findings—specifically, the shift in the populations of component and pattern cells from the V1 to the middle temporal cortex—and also the psychophysical findings such as adaptive global motion pooling. The utility of the attention mechanism during motion integration may be attributable to its similarity to the human visual grouping mechanism⁵¹.

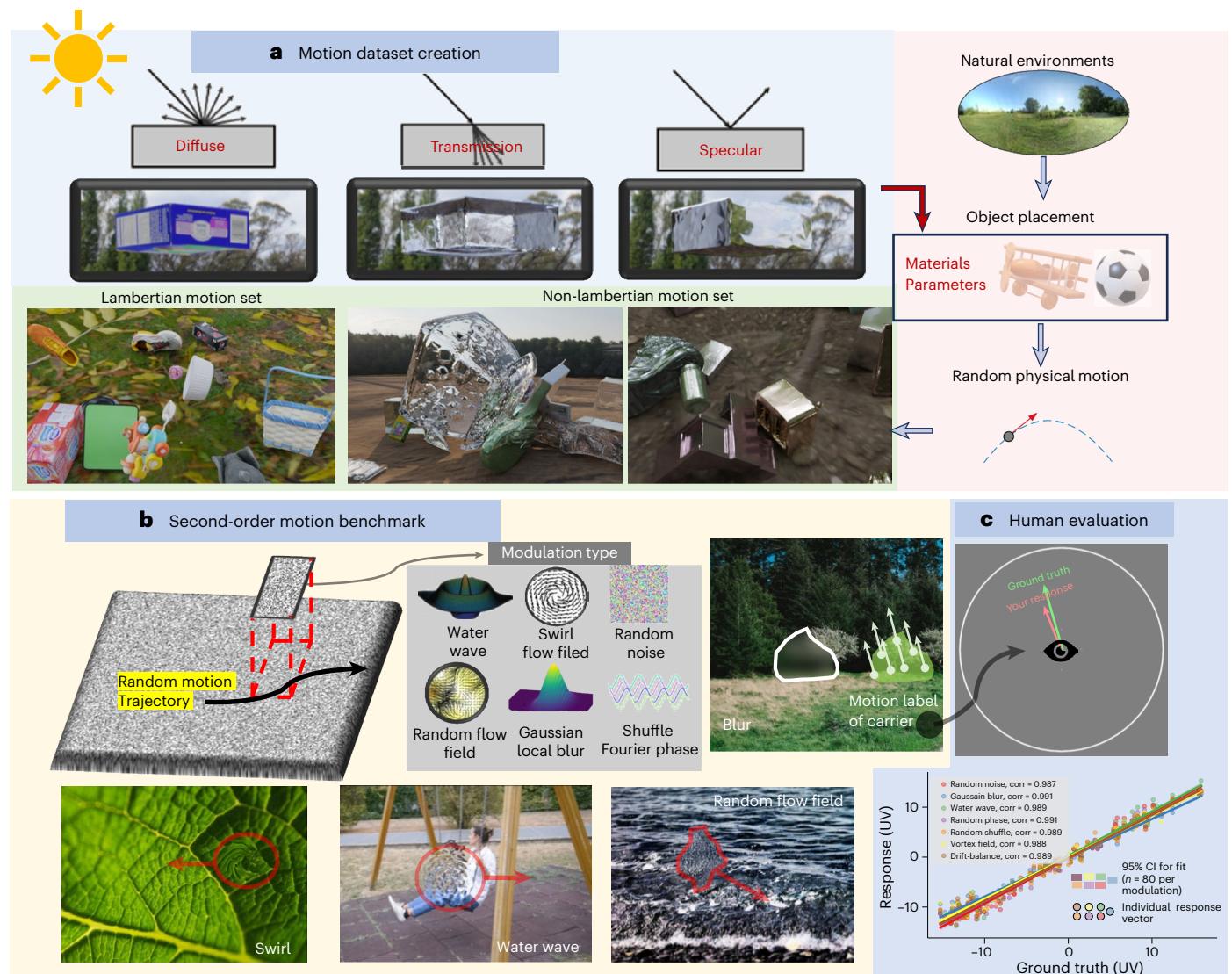


Fig. 4 | A material-controlled motion dataset and a second-order benchmark demonstration. a, We manipulated material properties to create two motion datasets with optical flow labels—one with purely reflective materials and another incorporating non-Lambertian surfaces such as specular, glossy, translucent and anisotropic materials. The motion was simulated by a physics engine with gravity and initial movement, whereas material properties were rendered via the Blender⁸⁷ engine. **b**, A large second-order benchmark was generated by applying naturalistic modulations—such as water waves and swirl effects—to natural images. A total of seven types of modulations were created

to evaluate both human and model responses. For illustrative purposes, the background images shown here have been replaced with visually similar, copyright-free alternatives. **c**, Psychophysical experiments using this dataset demonstrated that humans reliably perceive a wide range of second-order motions, whereas current machine vision models struggle with this task. The figure below illustrates perceived-motion vectors from a single participant across seven different modulations. The shaded region around the fitted line represents the 95% CI.

Second-order motion processing

Another critical contribution is that we reveal a function of second-order motion perception, which has received little attention from the CV community because the functional importance thereof has been poorly understood. Early studies suggested that visual analysis of second-order features might aid recognition of global image spatial structure⁵² and/or may distinguish separation by shading from a material change⁵³. However, the importance of second-order motion remained unclear. Here we show that biological systems may engage in second-order motion perception to ensure reliable motion estimation from non-diffuse material. This is an important advance in making CV algorithms more human-aligned and simultaneously more robust in estimating the dynamic structural changes of natural scenes. Our study also shows that machine learning can afford conceptual proof of neuroscientific hypotheses that suggest how specific functions evolved in natural environments.

Relationship with computer vision models

This study does not seek to outperform SOTA CV models optimized for certain engineering tasks. We instead employ a heuristic approach to balance the alignment of human vision with the robust processing of natural scenes. Inspired by the human visual system, it may be possible to expand the capacities of CV models. For example, we show that human-aligned computation efficiently captures inherent human-perceived flow illusions that CV models often fail to replicate (Table 1). The current CV methods, when presented with certain scenarios, are often unstable because they seek to match the pixel correspondences between frame pairs⁹. This strategy differs from the human higher-order motion perception mechanism, which depends on spatiotemporal features and demonstrates exceptional stability and adaptability in interpreting object motion. Furthermore, the second-order motion system could detect long-range motions of high-level features. The addition of this system not only combats noise

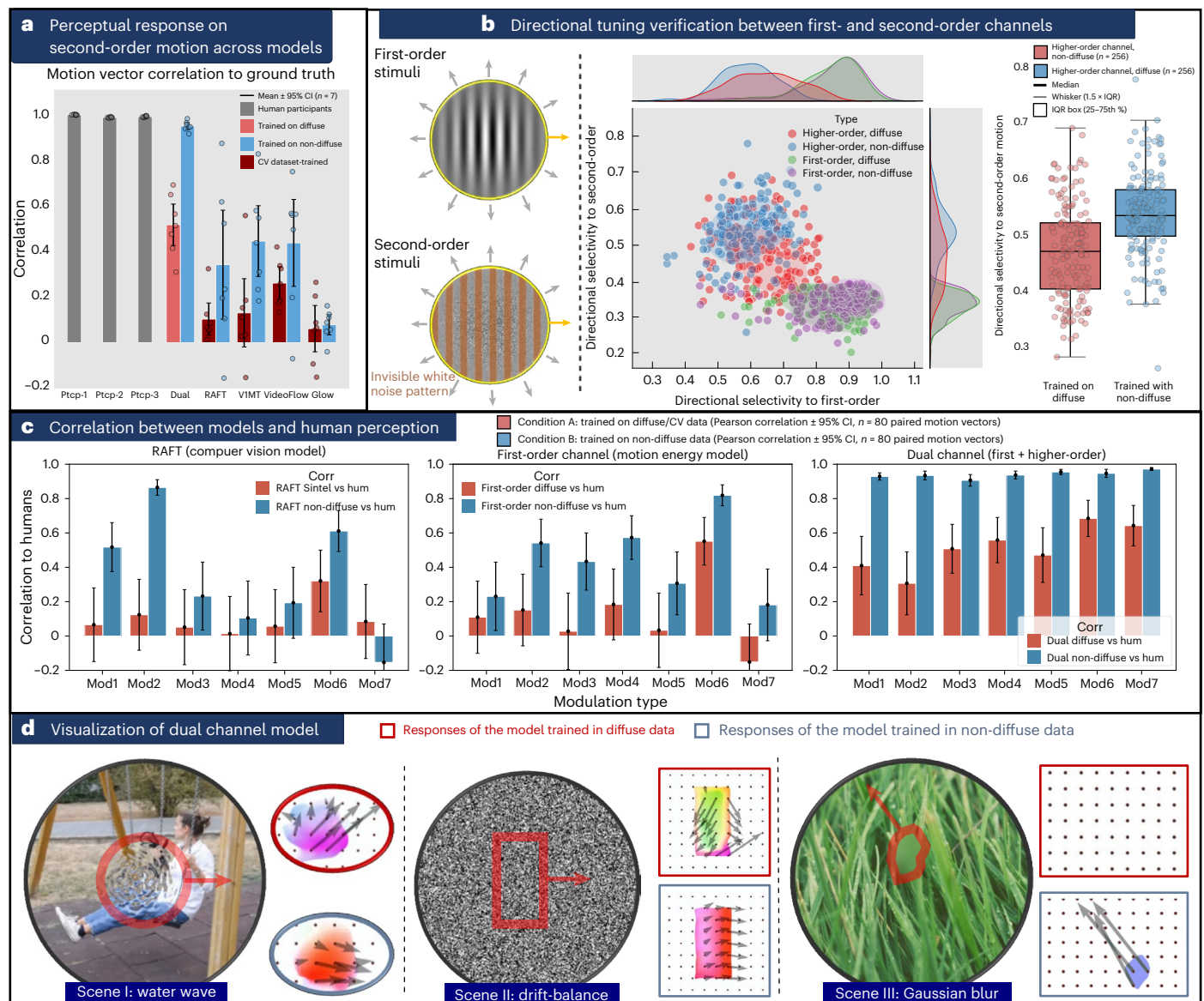


Fig. 5 | The interplay between material properties and second-order motion perception. **a**, We computed average Pearson correlations across various second-order motion types and compared them with contemporary CV models. Our dual-channel model, trained on a non-diffuse dataset, outperformed others—achieving near-human performance. Error bars indicate the 95% CI across seven modulation types; Ptcp-N represents different participants. **b**, The directional tuning curves of model units were evaluated using first- and second-order gratings. Tuning was quantified using a modified circular variance measure⁴², ranging from 0 (low tuning) to 1 (high tuning). Results from both first- and higher-order channels, trained on diffuse and non-diffuse datasets,

show that the higher-order channel exhibits significantly better tuning, further enhanced with non-diffuse data. IQR, interquartile range. **c**, Pearson correlations with human responses are detailed for all modulation types. Mod1 to Mod7 represent different second-order modulations, including random noise, Gaussian blur, water waves, Fourier phase shuffle, random pixel shuffle, swirl and drift-balanced motion. The dual-channel model trained on non-diffuse datasets demonstrated significantly improved recognition of second-order motion. Error bars denote the 95% CI for each modulation type. hum, human. **d**, The model's response to second-order motion stimuli trained separately on diffuse and non-diffuse motion sets. See Extended Data Tables 2 and 3 for quantitative data.

and optical turbulence but also results in a more stable and reliable motion estimation model, particularly useful in challenging scenarios such as adversarial attacks⁹ or extreme weather conditions⁵⁴. We believe these advancements offer substantial insights towards enhancing motion estimation in the CV field and developing a more reliable and stable model.

Limitations

Human-like visual systems require more than basic motion energy computation; they also need adaptive motion integration and higher-order motion feature extraction. Although our approach uses multilayer 3D

CNNs and motion graphs to address these needs, this inevitably reduces interpretability compared with more traditional models. Interpreting the specific higher-order features being extracted remains challenging, as does understanding how a dynamic graph structure could be implemented in real neural systems.

Although our dual-channel model simply integrates outputs from the two channels before the middle temporal cortex module, biological systems are known to adaptively use first- and higher-order channels depending on the stimulus condition (for example, jump size, retinal eccentricity and attention)^{26,55}. To mimic the adaptive switching, we manually switch off the higher-order channel when analysing the

phenomena in which first-order processing is supposed to dominate (refer to the ‘The two-stage processing model’ and ‘Motion graph-based scene integration’ sections). Even when switching higher-order channel on, we find no qualitative differences in the model prediction with regard to motion integration and illusions. For quantitative evaluation on naturalistic movie benchmarks, however, the addition of the higher-order channel reduces the response similarity to humans (Extended Data Table 4), presumably because the higher-order channel has a powerful 3D CNN that has no explicit human-aligned computational constraints. In the future, we would like to add a function to the model that can adaptively integrate dual-channel outputs in a way that is consistent with biological systems.

For the second-order motion benchmark, due to the technical challenges of real-world data collection, we use synthetic data for quantitative evaluation, acknowledging a potential gap between simulation and reality. Further data and validation would be helpful for practical applications in future work.

Finally, higher-order motion processing serves broader functions, including self-location and navigation in dynamic environments^{56–58}, and hierarchical decomposition of motion and object inference^{59,60}. These aspects are not explicitly modelled here; however, our model exhibits grouping and segmentation capacities based on motion inference, which are important steps toward hierarchical inferences of natural scenes.

Methods

Model structure

Our biologically oriented model features two stages, stages I and II. As shown in Fig. 1, stage I has two channels, of which the first engages in straightforward luminance-based motion energy computation, whereas the second contains a multilayer 3D CNN block that enables higher-order feature extraction.

Stage I (first-order channel). Spatiotemporally separable Gabor filter. When building our image-computable model, each input was a sequence of grayscale images $\mathbf{S}(\mathbf{p}, t)$ of spatial positions $\mathbf{p} = (x, y)$ within domain Ω at times $t > 0$. We sought to capture local motion energies at specific spatiotemporal frequencies, as do the direction-selective neurons of the V1 cortex. We modelled neuron responses using 3D Gabor filters^{61,62}. To enhance computational efficiency, these were decomposed into spatial 2D Gabor filters $\mathcal{G}(\cdot)$ and temporal 1D sinusoidal functions exhibiting exponential decay $\mathcal{T}(\cdot)$. Given the coordinates $x' = x \cos \theta + y \sin \theta$ and $y' = -x \sin \theta + y \cos \theta$, the filters may be defined as follows:

$$\begin{cases} \mathcal{G}(x, y; f_s, \theta, \sigma, \gamma) = \exp\left(-\frac{x'^2 + y'^2}{2\sigma^2}\right) \times e^{(2\pi f_s x')i}, \\ \mathcal{T}(t; f_t, \tau) = \exp\left(-\frac{t}{\tau}\right) \exp(2\pi i(f_t t)), \\ \text{s.t. } \{x, y, t \mid 0 \leq t < T; (x, y) \mid (x^2 + y^2 \leq R^2)\} \end{cases} \quad (1)$$

Trainable parameters such as f_s, f_t, θ, σ and γ control spatiotemporal tuning, orientation and the Gabor filter shape, whereas τ adjusts temporal impulse response decay. All parameters are subject to certain numerical constraints, for example, θ is limited to $[0, 2\pi)$ to avoid redundancy, whereas f_s and f_t are limited to less than 0.25 px per frame to avoid spectrum aliasing, and so on. The response L_n to the stimuli $\mathbf{S}(\mathbf{p}, t)$ is computed via separate convolutions:

$$\begin{aligned} L_n(x, y, t; \Theta) &= (\mathbf{S} * \mathcal{G}) * \mathcal{T} \\ &= \iiint \mathbf{S}(x, y, \mathcal{T}) \cdot \mathcal{G}_n(x - x', y - y') \\ &\quad \mathcal{T}_n(t - \mathcal{T}) dx dy d\mathcal{T} + \alpha_1 \end{aligned}$$

where α_1 are the learned spontaneous firing rates. Furthermore, local motion energy is captured by a phase-insensitive complex cell in the V1

cortex, which computes the squared summation of the response from a pair of simple V1 cells with orthogonal receptive fields⁶³, defined as (even and odd):

$$\begin{cases} L_n^o(x, y, t; \Theta) = \mathbf{S} * \Re[\mathcal{G}] * \Re[\mathcal{T}] + \mathbf{S} * \Im[\mathcal{G}] * \Im[\mathcal{T}] \\ L_n^e(x, y, t; \Theta) = \mathbf{S} * \Re[\mathcal{G}] * \Re[\mathcal{T}] - \mathbf{S} * \Im[\mathcal{G}] * \Im[\mathcal{T}], \end{cases} \quad (2)$$

where $\Re(\cdot)$ and $\Im(\cdot)$ extract the real and imaginary parts of a complex number and the asterisks denote convolution operations. The complex cell response L_n^c is then:

$$L_n^c(x, y, t; \Theta) = (L_n^o(x, y, t; \Theta))^2 + (L_n^e(x, y, t; \Theta))^2 \quad (3)$$

Multiscale wavelet processing. The convolution kernel of our spatial filter has a fixed size of 15×15 . This imposes a physical limitation on the receptive field of each unit. We employed a multiscale processing strategy to enhance receptive field size flexibility. Specifically, we constructed a pyramid of eight images that were linearly scaled from $H \times W$ to $\frac{H \times W}{16}$. The 256 complex cells are evenly distributed across the eight scales, with 32 cells per scale. All of these cells function as motion energy detectors, differing only in their receptive field sizes. Specifically, cells at coarser scales have larger receptive fields due to image downsampling before input. This enables the representation of different groups of cells that were sensitive to short- and long-distance motions⁶⁴. The $N = 256$ complex cells $\{L_n^c\}_i^N$ capture motion energy on multiple scales. We subjected each cell to energy normalization to ensure that the energy levels were consistent:

$$\hat{L}_n^c(t) = \frac{K_1 L_n^c(t)}{\sum_{i=1}^N L_i^c(t) + \sigma_1}, \quad (4)$$

where σ_1 is the semi-saturation constant of normalization and $K_1 > 0$ determines the maximum attainable response. We interpret the response, denoted $\hat{L}_n(t)$, as the model equivalent of a post-stimulus time histogram, which is a measure of the neuron's firing rate. Physiologically, such responses could also be computed using inhibitory feedback mechanisms^{65,66}. Bilinear interpolation was used to resize the multiscale motion energies to the same spatial size, thus $\frac{H \times W}{8}$. In the DNN context, this balances the trade-off between the spatial resolution and the computational overhead. The final output of the first stage is a 256-channel feature map $\mathbf{E}_1 \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 256}$ that captures the underlying, local motion energy and thus partially characterizes the cellular patterns of the V1 cortex in a computational manner⁶³; the implementation also illustrated in Extended Data Fig. 2a.

Stage I (higher-order channel). In the higher-order channel, we employ standard 3D CNNs to extract non-first-order features. This channel features five layers of 3D CNNs, each of kernel size $3 \times 3 \times 3$, linked via residual connections and nonlinear ReLU activation functions. The 3D CNN layers engage in preprocessing before extraction of nonlinear features, which are then processed using the motion energy constraints described above, and the motion energies calculated. As the human higher-order motion mechanism is highly sensitive to colour⁶⁷, each input to this channel is a sequence of RGB images, and the output is formatted to match that of the first-order channel: $\mathbf{E}_2 \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 256}$. Both the first- and higher-order channel activations undergo the same normalization process, after which they are merged via a 1×1 convolution. The resulting fused output $\mathbf{E}_m \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 256}$ is then fed to stage II.

In Fig. 5 and Extended Data Fig. 1, we designate the model incorporating stage II with $\mathbf{E}_m$ as **Ours-dual** (signal from the dual channel), whereas the model using only $\mathbf{E}_1$ is referred to as **Ours-first** (signal only from the first-order channel). To simplify discussions on motion-energy-based processing and integration (refer to the ‘The two-stage processing model’ and ‘Motion graph-based scene

integration' sections), we focus on the first-order channel, avoiding the complexities introduced by higher-order motion. Conversely, when analysing second-order motion perception (refer to the 'Material properties and second-order motion perception' and 'The interplay between the first- and higher-order channels' sections), we adopt the dual channel, jointly considering both first- and higher-order channels (Fig. 5 and Extended Data Fig. 1).

Stage II (global motion integration and segregation). First-stage neurons have a limited receptive field, constraining them to detect only nearby motion. Solving the aperture problem in motion-perception systems necessitates flexible spatial integration⁶⁸. This process involves complex mechanisms^{69,70} and requires extensive prior knowledge, which may surpass traditional modelling methods. Convolutional neural networks, with their extensive parameterization and adaptability, provide a viable solution; however, spatial integration of local motions demands more versatile connectivity than that offered by standard 3×3 convolutions, which are limited to local receptive fields. To address this, we developed a computational model that employed a graph network and recurrent processing for effective motion integration.

Motion graph based on a self-attention mechanism. We move beyond traditional Euclidean space in images, creating a more flexible connection across neurons using an undirected weighted graph, $\mathbf{G} = \{\mathbf{V}, \mathbf{A}\}$. Here, $\mathbf{V}$ denotes nodes (each spatial location $p(i, j)$) and $\mathbf{A}$ is the adjacency matrix, indicating connections among nodes. The feature of each node is the entire set of the corresponding local motion energies: $\mathbf{E}(i, j) \in \mathbb{R}^{1 \times 256}$. The connection between any pair of nodes is computed using a specific distance metric. Strong connections form between nodes with similar local motion energy patterns. This allows the model to establish connections flexibly between different moving objects or elements across spatial locations, thus creating what we term a motion graph. Specifically, the distance between any pair of nodes (i, j) is calculated using the cosine similarity. This is similar to the self-attention mechanisms of current transformer structures^{71–73}. We use the adjacency matrix $\mathbf{A} \in \mathbb{R}^{HW \times HW}$ to represent the connectivity of the whole topological space, where $\mathbf{A}$ is a symmetrical, semi-positive definite matrix defined as:

$$\mathbf{A}(i, j) = \mathbf{A}(j, i) = \frac{\varphi(\mathbf{E}_i) \cdot \varphi(\mathbf{E}_j)}{\|\varphi(\mathbf{E}_i)\| \|\varphi(\mathbf{E}_j)\|}. \quad (5)$$

We subject the connections between graphs to exponential scaling using the matrix $\mathbf{A}$ given by $\exp(\mathbf{A}s)$, where s is a learnable scalar restricted to within (0,10) to avoid overflow. The smaller the s , the smoother the connections across nodes, and vice versa. Finally, a symmetrical normalization operation balances the energy, resulting in $\mathbf{A} := \mathbf{D}^{-\frac{1}{2}} \exp(\mathbf{A}s) \mathbf{D}^{-\frac{1}{2}}$, where $\mathbf{D}$ is the degree matrix. This yields an energy-normalized undirected graph. Intuitively, the adjacency matrix represents the affinity or connectivity of a neuron within the space. Strong global connections form between neurons, the motion responses of which are related.

Recurrent integration processing. Recurrent neural networks flexibly model temporal dependencies and feedback loops, which are fundamental aspects of neural processing in the brain⁷⁴. We use a recurrent network, rather than multiple feedforward blocks, to simulate the process of local motion signals being gradually integrated into the middle temporal cortex and eventually converging to a stable state.

During each iteration i , an adjacency matrix $\mathbf{A}^i$ is first constructed using the current graph embedding feature $\mathbf{A}^i$. Subsequent motion integration is achieved through a simple matrix multiplication. We introduce the gated recurrent unit⁷⁵, implemented in a convolutional manner³⁹, as a general component for propagating memory from the

current state to the next iteration. The integrated motion information is therefore passed through convolutional gated recurrent unit blocks that update the motion energies:

$$\mathbf{E}^{i+1} = \text{GRU}_\theta(\mathbf{A}^i \times \mathbf{E}^i, \mathbf{E}^i) \quad (6)$$

This is computationally similar to the information propagation mechanisms in transformers^{71,72} and can also be viewed as a simplified form of graph convolution⁷⁶. Through recurrent iteration, this motion integration approximates the ideal final convergence of motion energies, that is, $\mathbf{E}_k \rightarrow \mathbf{E}^*$.

We adopted the same approach to decode the 2D optical flow from $\mathbf{E}$ of each iteration k . Specifically, the integrated motion $\mathbf{E}$ is squared to ensure positivity and then normalized in terms of energy:

$$\hat{\mathbf{E}}(i, j) = K_2 \mathbf{E}^2(i, j) / \sum_{i, j} \mathbf{E}^2(i, j) + \sigma_2^2.$$

This yields $\hat{\mathbf{E}} \in \mathbb{R}^{H \times W \times 256}$, which could be viewed as a post-stimulus time histogram of neuronal activation. We use a shared flow decoder to project the activation pattern of each spatial location onto the motion field $F \in \mathbb{R}^{H \times W \times 2}$. This decoder employs multiple 1×1 convolution blocks with residual connections, as do recent advanced optical flow models^{77,78}. We observed that the results generally converged by the eighth iteration. This was therefore chosen as the standard stage II output. The overall inference pipeline is illustrated in Extended Data Fig. 2.

Cutting of an object instance from the motion graph. The interactions of objects in a dynamic scene are reflected in the adjacency matrix of the motion graph $\mathbf{G}$. After the incorporation of this adjacency matrix into $\mathbf{A} \in \mathbb{R}^{HW \times HW}$, segmentation can be achieved using a graph-cut method. Specifically, we employ the normalized cuts (Ncut) method¹⁵. This partitions a graph into disjoint subsets by minimizing the total edge weight between the subsets relative to the total edge weight within each subset. Specifically, the Laplacian matrix of $\mathbf{G}$ can be expressed as $\mathbf{L} = \mathbf{D} - \mathbf{A}$, or in the symmetrically normalized form as $\mathbf{L} = \mathbf{I}_n - \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}}$, where $\mathbf{D}$ is a diagonal matrix defined as $\mathbf{D} = \text{diag}(\{\sum_j A_{ij}\}_{i=1}^n)$; $\mathbf{L}$ is a semi-positive definite matrix, which facilitates the orthogonal decomposition to yield $\mathbf{L} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T$, where $\mathbf{U}$ is the set of all orthonormal basis vectors, denoted as $\{u_i\}_{i=1}^n$ and is therefore the Fourier basis of $\mathbf{G}$. The $\mathbf{\Lambda}$ term is a diagonal matrix containing all eigenvalues $\{\lambda_i\}_{i=1}^n$ ordered as $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$. According to ref. 15, the eigenvector corresponding to the second smallest eigenvalue, $u_2 \in \mathbb{R}^{HW}$, commonly termed the Fiedler vector, yields a real-valued solution to the relaxed Ncut problem. In our implementation, we extract u_2 and then apply binarization using the rule $u_2 = u_2 > \text{mean}(u_2)$. The resulting binary segmentation is viewed as a potential field and further refined using a conditional random field⁷⁹. As such binarization does not inherently distinguish between foreground and background, we adaptively assign a polarity that matches the foreground during evaluation using the DAVIS 2016 segmentation benchmark. The results shown in the second row of Extended Data Fig. 1 were obtained using a recurrent bipartitioning method⁸⁰ that allows multi-object segmentation. Notably, the entire process is training free.

Training strategy

We employ a supervised learning approach to minimize the difference between the model's predictions and physical ground truth, and human motion perception data is only used for evaluation. Our primary focus is on how effectively the model mimics human motion perception, rather than how precisely it predicts the ground truth. During training, we use a sequential pixel-wise mean-squared-error loss to minimize the difference between the ground truth and the model predictions of stage I (and of each iteration of stage II).

Dataset. Our dataset encompasses a diverse range of natural and artificial motion scenes. Specifically, it integrates existing benchmarks such as MPI-Sintel, Sintel slow²⁸ and KITTI^{29,30}, along with natural videos from DAVIS, where pseudo-labels are generated using FlowFormer⁸¹. This collection is referred to as dataset A.

We also introduce custom multi-frame datasets: dataset B, which comprises simple non-textured 2D motion patterns, and dataset C, which features drifting grating motions (that is, continuously translating sinusoidal gratings with orthogonal ground-truth motion directions). These datasets provide fundamental motion patterns that facilitate training from scratch, accelerating convergence and improving model stability³⁸. Furthermore, as suggested by ref. 8, incorporating such datasets aids in model adaptation to non-textured scenarios and introduces an orthogonal motion bias to ambiguous motion. It remains controversial whether this bias reflects a slow-world Bayesian prior⁸² or other causes¹⁰.

To study second-order motion, we developed datasets with diffuse (dataset D) and non-diffuse (dataset E) objects and integrated them into training. We then evaluated how the model perceived material properties and second-order motion. We define three training types:

1. **Types I and II:** The model was trained separately on D and E to assess how second-order motion perception is related to material properties. Results for these models, referred to as **Ours-D** (diffuse) and **Ours-ND** (non-diffuse), are shown in Fig. 5b,c.
2. **Type III:** The model was trained on a mixed dataset {A, B, C, D, E} using a curriculum strategy, thus starting with {B, C} and progressing to the full set. This approach, commonly used during optical flow model training^{38,39}, improves convergence and robustness. All of the other results are based on type III training, denoted by the **Ours-F** (final). Unless otherwise specified, **Ours** refers to **Ours-F** throughout all of the results.

The environment. Model training was performed in PyTorch 2.0 on a workstation equipped with five NVIDIA RTX A6000 GPUs operating in parallel under the CUDA v.11.7 runtime. Human psychophysical data were collected using Python v.3.9.12 alongside PsychoPy v.2023.2, EasyDict v.1.10, Pandas v.2.0.0 and NumPy v.1.23.5.

Data analysis and visualization were performed in MATLAB v.2023a and Python v.3.9.12 by using NumPy v.1.23.5, Pandas v.2.0.0, Matplotlib v.3.7.5, Seaborn v.0.13.2, SciPy v.1.7.3 and Pingouin v.0.5.3. All code is available at ref. 83.

Timing. Given the standard playback frame rate of 25 fps and the human visual impulse response duration of approximately 200 ms, we configured the temporal window of stage I to cover six frames (200 ms). For the first-order channel, sequences of 11 consecutive greyscale images were input. Supervised training uses the instantaneous velocity at the sequence midpoint (that is, the fifth frame) as the training label. The higher-order channel with the 3D CNN was trained using a longer temporal sequence of 15 frames to capture long-term spatio-temporal features effectively.

Dataset generation

Simple motion generation. To generate simple motion in dataset B, we employ an image-based affine transformation to warp objects and simulate various motion patterns. Specifically, we first create multiple sub-regions with different shapes (for example, circles, rectangles or super-pixel partitions⁸⁴) atop a background of uniform random colours. We then select n sub-regions as moving elements and place them randomly in the first image.

We simulate multi-frame motion under the assumption that object motion remains smooth, as is the case in natural environments. To this

end, we partially adopt a Markov chain principle, where an object's motion state $\mathbf{S}(t) = [U(t), V(t)]$ depends only on $\mathbf{S}(t-1)$:

$$\begin{aligned}\Pr[\mathbf{S}(t) = s_t | \mathbf{S}(t-1) = s_{t-1}, \dots] \\ = \Pr[\mathbf{S}(t) = s_t | \mathbf{S}(t-1) = s_{t-1}].\end{aligned}\quad (7)$$

The motion state at time t follows a 2D Gaussian:

$$[U(t), V(t)] \approx \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_U^2 & 0 \\ 0 & \sigma_V^2 \end{pmatrix}, \quad (8)$$

where $\boldsymbol{\mu} = [U(t-1), V(t-1)]^T$. We set σ_U, σ_V as constants controlling motion variability, ensuring random yet smooth motion for each object. The initial state $\mathbf{S}(0)$ is similarly random, with speed $|\mathbf{S}(0)|$ drawn from $\mathcal{N}(\mu, \sigma)$ and angle from a uniform distribution $\mathcal{U}(0, 2\pi)$. The parameters μ and $\sigma = \frac{\mu}{3}$ are chosen to match empirical speed distributions in the training set.

In practice, we simulate translation, rotation, scaling and distortion for each element. These transformations all obey the proposed Markov process to preserve smooth motion. At each time step, we apply sequential affine transformations on a uniform 2D grid using PyTorch's `affine_grid` and `grid_sample` for GPU acceleration. The optical flow ground truth is derived via the inverse of these transformations.

Dataset rendering. To generate datasets D and E (Fig. 4a), we used the Kubric pipeline⁸⁵ to synthesize large-motion datasets that integrate PyBullet⁸⁶ for physics simulation and Blender⁸⁷ for photorealistic rendering. A variety of 3D models and textures were selected from ShapeNet and GSO, whereas natural HDRI backgrounds from Polyhaven⁸⁸ provided realistic illumination. For the diffuse (Lambertian) motion dataset, we generated 58 scenes with a static camera and 35 scenes with dynamic camera motion. By contrast, the non-diffuse (non-Lambertian) dataset comprises 131 static scenes and 27 scenes with dynamic camera motion. Each scene consists of 36 consecutive frames rendered at a resolution of 768×768 px, 30 fps. Scene composition was carefully controlled through a series of configurable parameters. In each scene, the number of static (distractor) objects was randomly chosen between 7 and 15, whereas the number of dynamic (tossed) objects ranged from 5 to 12. Static objects were spawned within a predefined region bounded by the coordinates $(-7, -7, 0)$ and $(7, 7, 10)$ (in metres), whereas dynamic objects were placed in a more restricted region between $(-5, -5, 1)$ and $(5, 5, 5)$. Their initial velocities were uniformly sampled from the range $[(-2, -2, 0), (2, 2, 0)]$, which ensured diverse motion trajectories under controlled friction and restitution conditions. Camera configurations were designed to capture different motion types. In the fixed configuration, the camera was randomly positioned within a half-spherical shell and aimed at the scene centre. For dynamic acquisition, the camera underwent linear motion by interpolating between two independently sampled positions, with the maximum displacement limited to 4 m s^{-1} . Optical flow labels were automatically generated using Kubric's built-in functions, which track the displacement of each element in camera coordinates and project these displacements into pixel coordinates.

Material properties were manipulated via the principled BSDF function to achieve natural optical effects. Materials with Lambertian reflectance were employed for diffuse scenes, whereas non-diffuse scenes featured materials with increased metallicity, specular, anisotropy and transmission. In the latter settings, the material assignment was randomized from a set of predefined functions (for example, those assigning metallic, anisotropic or transmission properties) to yield a varied yet natural appearance across objects. All other aspects—such as illumination, object placement, and scene configuration—were standardized across datasets to ensure consistency.

Second-order motion modulation. As illustrated by Fig. 4b, we developed a second-order dataset to benchmark perception capability in both humans and computational models. The dataset consists of 40 scenes featuring seven types of second-order motion modulations. Each modulation comprises 16 frames, with a randomly moving carrier overlaid on a 1,024 × 1,024 natural image background selected from an open-sourced image dataset⁸⁹. To eliminate first-order motion interference, the natural images were kept static and the random motion patterns were generated using a similar Markov chain from equation (7), where the motion states $[U, V]$ were sampled from 2D Gaussian distributions conditioned on the previous state. The carrier was subjected to seven distinct second-order motion modulations, encompassing spatial effects such as {Gaussian blurring}; temporal effects such as {drift-balanced} motion and {shuffle Fourier phase}; and spatiotemporal effects such as {water waves} and {swirls}. The spatial noise and blur were sparse Gaussian noise and localized Gaussian blur, respectively. The water wave, swirl and random flow field modulations warp pixels using specific flow fields. In terms of the water wave dynamics, the flow field $f_{u,v,t} = [\frac{\partial K}{\partial x}, \frac{\partial K}{\partial y}]$ was:

$$K(r, t) = \cos(2\pi f r) \times e^{-\nu^2} \times \cos(2\pi \xi t) \times e^{-\delta t^2},$$

$$r = \sqrt{x^2 + y^2},$$

where f , ξ and δ control the wave frequency, temporal variation and damping, respectively. We superimposed multiple water waves that differed in terms of their dynamics in different locations. This created chaotic, local optical turbulence contemporaneous with carrier motion. The real carrier motion was thus obscured by local optical noise and was invisible in Fourier space, epitomizing the characteristics of second-order motion. Similarly, {random flow field} or {shuffle Fourier phase} modulation involves the warping of either the pixels or the Fourier phase of original local regions using a randomly sampled Gaussian flow field.

Experimental details

In silico neurophysiological methods. We employed drifting Gabor or plaid (composed of two Gabor components) with a single frequency component as the input stimulus. For second-order motion, drift-balanced motion modulation was applied to the same Gabor envelope.

The model responses after stage I and after each iteration of stage II were considered analogous to the post-stimulus time histogram of a neuron, thus reflecting activation levels. Responses across the spatial dimensions were averaged to obtain the activation distributions of the 256 units, represented as $\mathbb{R}^{1 \times 1 \times 256}$ with respect to the input stimulus. The stimuli were typically 512 × 512 px in size, with full contrast.

Directional tuning. We employed a single frequency drifting Gabor and a plaid (superimposed at $\pm 30^\circ$) as stimuli. Initially, twelve directions were uniformly sampled from $(0, 2\pi]$. For each direction, we logarithmically sampled $8 \times 8 = 64$ sets of spatiotemporal frequency combinations and used the drifting Gabor stimulus to obtain 64 directional tuning curves for each unit. The spatiotemporal frequency with the largest standard deviation was selected as the preferred frequency st^* for each unit. Gabor and plaid stimuli with the frequency configurations of st^* were then input to the model to derive the directional tuning curves of all units. The model tuning curve with st^* as the drifting Gabor was termed $\mathcal{C}$ and that for the plaid $\mathcal{P}$. We next assessed the directional tuning capacity by deriving partial correlations³²:

$$\begin{cases} R_{\text{pattern}} = \frac{r_p - r_c r_{cp}}{\sqrt{(1-r_c^2)(1-r_{cp}^2)}}, \\ R_{\text{component}} = \frac{r_c - r_p r_{cp}}{\sqrt{(1-r_p^2)(1-r_{cp}^2)}}, \end{cases} \quad (9)$$

where r_c is the correlation between $\mathcal{P}$ and the component prediction that is the superimposed $\pm 30^\circ$ shift of $\mathcal{C}$; r_p is the correlation between $\mathcal{P}$ and the pattern prediction $\mathcal{C}$; r_{cp} is the correlation between these two predictions. Units were classified as component, pattern or unclassified on the basis of these correlations (Fig. 2a).

Orientation selectivity quantification. Figure 5b shows how the orientation selectivity O_{ori} was quantified using the modified circular variance⁴²:

$$O_{\text{ori}} = \left| \frac{\sum_i A(\theta_i) \exp(2i\theta_i)}{\sum_i A(\theta_i)} \right|, \quad (10)$$

where $A(\theta_i)$ is the normalized response at angle θ_i .

Human and model comparison. We used the human-perceived flow data⁷ of the Sintel and KITTI 2015³⁰ benchmark for comparison. The metrics include the vector endpoint error, the Pearson correlation and the partial correlation. Partial correlation measures the relationship between human responses and model predictions after controlling for the ground truth:

$$r_{\text{resp model-GT}} = \frac{r_{\text{resp model}} - r_{\text{resp GT}} \times r_{\text{model GT}}}{\sqrt{1-r_{\text{resp GT}}^2} \sqrt{1-r_{\text{model GT}}^2}}, \quad (11)$$

where r is the Pearson correlation. In addition, the RCI is an index from ref. 7 to evaluate the similarity between model performance and human flow illusions at each probed location.

The RCI is defined as the product of A·B·C in equation (14), measuring the relative alignment of ground truth (G), human response (R), model prediction (M) and the origin (O):

- A quantifies the deviation of human responses from the ground truth.
- B indicates the directional similarity between the response error vector $\vec{GR}$ and the model error vector $\vec{GM}$ relative to the ground truth.
- C compares the distance between model prediction and ground truth $\|\vec{GM}\|$ with the distance between model prediction and response $\|\vec{RM}\|$.

The RCI approaches +1 when the model's prediction aligns closely with human flow illusions and approaches -1 when the prediction diverges in the opposite direction.

$$A = \frac{\|\vec{GR}\|}{\|\vec{OG}\| + \|\vec{OR}\|}, \quad (12)$$

$$B = \frac{\vec{GR} \cdot \vec{GM}}{\|\vec{GR}\| \|\vec{GM}\|}, \quad (13)$$

$$C = 0.5 \left(\frac{\|\vec{GM}\| - \|\vec{RM}\|}{\|\vec{GM}\| + \|\vec{RM}\|} + 1 \right) = \frac{\|\vec{GM}\|}{\|\vec{GM}\| + \|\vec{RM}\|}. \quad (14)$$

Human data collection. We compared our model prediction with the human-perceived motions for the Sintel slow²⁸ benchmark (Table 1) using the data reported in ref. 7. The human data were collected in the laboratory with a strict yet practical psychophysical procedure. Briefly, in each trial, participants viewed repeated alternating presentations of the target motion sequence and a matching stimulus (Brownian noise). The spatiotemporal position of the target was indicated by a flash probe. Participants then used a mouse to adjust the speed and direction of the noise motion until it matched their subjective perception

of the target's motion. We recorded the matched noise motion as the participant's report of the subjective target motion.

The experiment controlled visual presentation across both spatial and temporal domains. Spatially, the display resolution is set at 50 px per 1° of visual angle. Temporally, visual stimuli were presented at 60 Hz for Sintel slow 4K resolution image sequences. To minimize directional bias, we applied data augmentation by flipping images horizontally and vertically, generating four replicated collections per data location. These flipped versions were averaged to mitigate orientation-dependent perceptual biases. Finally, each data point was averaged across 16 trials to ensure measurement stability.

To validate data reliability⁷, conducted a preliminary random dot kinematogram task to train and verify participants' performance before the main experiment. In this random dot kinematogram task, participants estimated the basic motion pattern of 5,000 black-and-white dots moving uniformly within a 600 px circular aperture. The results showed a strong, though not perfect, agreement between the reported motion and the ground truth motion (correlation = 0.97 in (u, v)), as illustrated in Figure 2 of ref. 7. As the target and matching stimuli were similar noise patterns in this task, it was relatively straightforward. These results indicate that our procedure can provide highly accurate estimates of human-perceived-motion vectors under optimal conditions. Data from MPI-Sintel (refer to Supplementary Figure 4 in ref. 7) further demonstrate that participants can accurately align the flash-probing in both space and time, yielding minimal endpoint errors relative to ground-truth vectors in neighbouring locations and time steps.

The human data for the KITTI 2025 benchmark was measured in an online experiment using a similar psychophysical method⁹⁰.

Second-order motion benchmark. We extended the paradigm in ref. 7 to collect second-order motion data. Stimuli were displayed on a VIEWPixx/3D LCD monitor (VPixx Technologies) with a resolution of 1,920 × 1,080 px at a 30 Hz refresh rate. The display luminance levels were linearly calibrated using an iPro chromometer (VPixx Technologies). The minimum, mean and maximum values were 1.8, 48.4 and 96.7 cd m⁻², respectively. The viewing distance was 70 cm and each pixel subtended 1.2376 arcmin. Participants sat in a darkened room using a chinrest to stabilize the head and performed experiments.

In each trial, a 600 px aperture at the screen centre displayed second-order motion for 500 ms (15 frames), followed by a 750 ms inter-stimulus interval, then 500 ms (15 frames) of brown noise within a 120 px aperture. A 15 px probe indicated the timing and location of the target motion, and four 5 px dots—orthogonally arranged 60 px from the display centre—served as position markers. During repeated presentations of the target motion and noise motion, participants used a mouse to adjust the noise motion's speed and direction until it matched their perception of the target's second-order motion, as illustrated in Extended Data Fig. 3. As the reported noise motion reflected the perceived target motion, it was recorded as the reported second-order perception. Seven types of second-order modulations were tested, each across 40 scenes. To counteract directional bias, each scene was presented in four variations—original, horizontally flipped, vertically flipped and both flipped—yielding 1,120 trials per participant over 6 h. Results were averaged across flipped versions into 280 perceived-motion vectors, which were then compared against computer vision models. The stimulus sequence was randomized for each participant.

The experiment adhered to the ethical standards of the Declaration of Helsinki, with the exception of preregistration, and was approved by the Ethics Committee of Kyoto University (approval no. KUIS-EAR-2020-003). Two authors and one naive participant (three males, average age 25.3 years) with normal or corrected-to-normal vision participated. Informed consent was obtained prior to the experiment. All participants were later financially compensated.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The project website is publicly available at <https://anonymized.github.io/motion-model-website/>. Human psychophysical data and the corresponding model responses are available at <https://github.com/anonymized/multi-order-motion-model> and are also archived on Zenodo⁸³. All other relevant data supporting the findings of this study—including model predictions, human behavioural responses and custom datasets (Drifting Grating, Non-textured 2D Motion, Diffuse Motion, Non-diffuse Motion and Second-order Motion datasets)—are provided at the same repository. Two additional mini motion datasets featuring diffuse and non-diffuse objects have also been made available to support quick verification of the effects on second-order motion perception. The public datasets used in this study are accessible from the following sources: Kubric, <https://github.com/google-research/kubric>; KITTI, <https://www.cvlibs.net/datasets/kitti/>; MPI-Sintel, <http://sintel.is.tue.mpg.de/>; Sintel-slow, https://www.cvlibs.net/projects/slow_flow/; DAVIS, <https://davischallenge.org/>; and Unsplash, <https://github.com/unsplash/datasets>.

Code availability

Our model implementation and human experimental code are publicly available at <https://github.com/anonymized/multi-order-motion-model>. This code can be accessed via <https://doi.org/10.5281/zenodo.14958959> (ref. 83). The code is released under the Apache License v.2.0.

References

1. Yamins, D. L. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
2. Kriegeskorte, N. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* **1**, 417–446 (2015).
3. Wichmann, F. A. & Geirhos, R. Are deep neural networks adequate behavioral models of human visual perception? *Annu. Rev. Vis. Sci.* **9**, 501–524 (2023).
4. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (eds Huttenlocher, D. et al.) 248–255 (IEEE, 2009).
5. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (eds Grauman, K. et al.) 3431–3440 (IEEE, 2015).
6. Dosovitskiy, A. et al. FlowNet: learning optical flow with convolutional networks. In *Proc. IEEE International Conference on Computer Vision (ICCV)* (eds Bajcsy, R. et al.) 2758–2766 (IEEE, 2015).
7. Yang, Y.-H., Fukiage, T., Sun, Z. & Nishida, S. Psychophysical measurement of perceived motion flow of naturalistic scenes. *iScience* **26**, 108307 (2023).
8. Sun, Z., Chen, Y.-J., Yang, Y.-H. & Nishida, S. Comparative analysis of visual motion perception: computer vision models versus human vision. In *Proc. Conference on Cognitive Computational Neuroscience* (eds Isik, L. et al.) 991–994 (CCN, 2023).
9. Ranjan, A., Janai, J., Geiger, A. & Black, M. J. Attacking optical flow. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* (eds Lee, K. M. et al.) 2405–2413 (IEEE, 2019).
10. Rideaux, R. & Welchman, A. E. But still it moves: static image statistics underlie how we see motion. *J. Neurosci.* **40**, 2538–2552 (2020).
11. Mineault, P., Bakhtiari, S., Richards, B. & Pack, C. Your head is there to move you around: goal-driven models of the primate dorsal pathway. In *Advances in Neural Information Processing Systems* (eds Ranzato, M. et al.) 28757–28771 (NeurIPS, 2021).

12. Nakamura, D. & Gomi, H. Decoding self-motion from visual image sequence predicts distinctive features of reflexive motor responses to visual motion. *Neural Netw.* **162**, 516–530 (2023).
13. Simoncelli, E. P. & Heeger, D. J. A model of neuronal responses in visual area MT. *Vis. Res.* **38**, 743–761 (1998).
14. Nishimoto, S. & Gallant, J. L. A three-dimensional spatiotemporal receptive field model explains responses of area MT neurons to naturalistic movies. *J. Neurosci.* **31**, 14551–14564 (2011).
15. Shi, J. & Malik, J. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 888–905 (2000).
16. Sun, Z., Chen, Y.-J., Yang, Y.-H. & Nishida, S. Modeling human visual motion processing with trainable motion energy sensing and a self-attention network. In *Advances in Neural Information Processing Systems* (eds Oh, A. et al.) 24335–24348 (NeurIPS, 2023).
17. Chubb, C. & Sperling, G. Drift-balanced random stimuli: a general basis for studying non-Fourier motion perception. *J. Opt. Soc. Am. A* **5**, 1986–2007 (1988).
18. Cavanagh, P. & Mather, G. Motion: the long and short of it. *Spat. Vis.* **4**, 103–129 (1989).
19. O’Keefe, L. P. & Movshon, J. A. Processing of first-and second-order motion signals by neurons in area MT of the macaque monkey. *Vis. Neurosci.* **15**, 305–317 (1998).
20. Theobald, J. C., Duistermars, B. J., Ringach, D. L. & Frye, M. A. Flies see second-order motion. *Curr. Biol.* **18**, R464–R465 (2008).
21. Baker, C. L. Jr Central neural mechanisms for detecting second-order motion. *Curr. Opin. Neurobiol.* **9**, 461–466 (1999).
22. Fleet, D. & Weiss, Y. *Optical Flow Estimation* (Springer, 2006).
23. Clifford, C. W., Freedman, J. N. & Vaina, L. M. First-and second-order motion perception in Gabor micropattern stimuli: psychophysics and computational modelling. *Cogn. Brain Res.* **6**, 263–271 (1998).
24. Ledgeway, T. & Smith, A. T. Evidence for separate motion-detecting mechanisms for first-and second-order motion in human vision. *Vis. Res.* **34**, 2727–2740 (1994).
25. Smith, A. T., Greenlee, M. W., Singh, K. D., Kraemer, F. M. & Hennig, J. The processing of first-and second-order motion in human visual cortex assessed by functional magnetic resonance imaging (fMRI). *J. Neurosci.* **18**, 3816–3830 (1998).
26. Nishida, S. & Ashida, H. A hierarchical structure of motion system revealed by interocular transfer of flicker motion aftereffects. *Vis. Res.* **40**, 265–278 (2000).
27. Prins, N. et al. Mechanism independence for texture-modulation detection is consistent with a filter-rectify-filter mechanism. *Vis. Neurosci.* **20**, 65–76 (2003).
28. Butler, D. J., Wulff, J., Stanley, G. B. & Black, M. J. A naturalistic open source movie for optical flow evaluation. In *Proc. Computer Vision—ECCV2012: 12th European Conference on Computer Vision* (eds Fitzgibbon, A. et al.) 611–625 (Springer, 2012).
29. Geiger, A., Lenz, P. & Urtasun, R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (eds Chellappa, R. et al.) 3354–3361 (IEEE, 2012).
30. Menze, M. & Geiger, A. Object scene flow for autonomous vehicles. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (eds Bischof, H. et al.) 3061–3070 (IEEE, 2015).
31. Fennema, C. L. & Thompson, W. B. Velocity determination in scenes containing several moving objects. *Comput. Gr. Image Process.* **9**, 301–315 (1979).
32. Movshon, J. A., Adelson, E. H., Gizzi, M. S. & Newsome, W. T. in *Pattern Recognition Mechanisms* (eds Chagas, C. et al.) 117–151 (Vatican, 1985).
33. Amano, K., Edwards, M., Badcock, D. R. & Nishida, S. Adaptive pooling of visual motion signals by the human visual system revealed with a novel multi-element stimulus. *J. Vis.* **9**, 4 (2009).
34. McDermott, J., Weiss, Y. & Adelson, E. H. Beyond junctions: nonlocal form constraints on motion interpretation. *Perception* **30**, 905–923 (2001).
35. Handa, T. & Mikami, A. Neuronal correlates of motion-defined shape perception in primate dorsal and ventral streams. *Eur. J. Neurosci.* **48**, 3171–3185 (2018).
36. Farneback, G. Two-frame motion estimation based on polynomial expansion. In *Proc. Image Analysis: 13th Scandinavian Conference (SCIA 2003)* (eds Bigun, J. & Gustavsson, T.) 363–370 (Springer, 2003).
37. Solari, F., Chessa, M., Medathati, N. K. & Kornprobst, P. What can we expect from a V1-MT feedforward architecture for optical flow estimation? *Signal Process. Image Commun.* **39**, 342–354 (2015).
38. Ilg, E. et al. FlowNet 2.0: evolution of optical flow estimation with deep networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (eds Chellappa, R. et al.) 2462–2470 (IEEE, 2017).
39. Teed, Z. & Deng, J. RAFT: recurrent all-pairs field transforms for optical flow. In *Proc. Computer Vision—ECCV2020 : 16th European Conference on Computer Vision* (eds Vedaldi, A. et al.) 402–419 (Springer, 2020).
40. Luo, A. et al. Learning optical flow with adaptive graph reasoning. In *Proc. 36th AAAI Conference on Artificial Intelligence* (eds Sycara, K. et al.) 1890–1898 (AAAI, 2022).
41. Xu, H., Zhang, J., Cai, J., Rezaatoughi, H. & Tao, D. GMFlow: learning optical flow via global matching. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (eds Dana, K. et al.) 8121–8130 (IEEE, 2022).
42. Mazurek, M., Kager, M. & Van Hooser, S. D. Robust quantification of orientation selectivity and direction selectivity. *Front. Neural Circuits* **8**, 92 (2014).
43. Shi, X. et al. VideoFlow: exploiting temporal cues for multi-frame optical flow estimation. In *Proc. International Conference on Computer Vision (ICCV)* (eds Agapito, L. et al.) 12469–12480 (IEEE, 2023).
44. Lucas, B. D. & Kanade, T. An iterative image registration technique with an application to stereo vision. In *Proc. 7th International Joint Conference on Artificial Intelligence (IJCAI ’81)* (ed. Hayes, P. J.) 674–679 (ACM, 1981).
45. Jaegle, A. et al. Perceiver IO: a general architecture for structured inputs & outputs. In *International Conference on Learning Representations (ICLR)* (eds Flinn, C. et al.) (2022).
46. Perazzi, F. et al. A benchmark dataset and evaluation methodology for video object segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (eds Agapito, L. et al.) 724–732 (IEEE, 2016).
47. Cheng, B., Misra, I., Schwing, A. G., Kirillov, A. & Girdhar, R. Masked-attention mask transformer for universal image segmentation. In *Proc. IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)* (eds Dana, K. et al.) 1290–1299 (IEEE, 2022).
48. Heo, M., Hwang, S., Oh, S. W., Lee, J.-Y. & Kim, S. J. VITA: video instance segmentation via object token association. In *Advances in Neural Information Processing Systems 35* (eds Koyejo, S. et al.) 23109–23120 (NeurIPS, 2022).
49. Rideaux, R. & Welchman, A. E. Exploring and explaining properties of motion processing in biological brains using a neural network. *J. Vis.* **21**, 11 (2021).
50. Storrs, K., Kampman, O., Rideaux, R., Maiello, G. & Fleming, R. Properties of V1 and MT motion tuning emerge from unsupervised predictive learning. *J. Vis.* **22**, 4415 (2022).
51. Mehrani, P. & Tsotsos, J. K. Self-attention in vision transformers performs perceptual grouping, not attention. *Front. Comput. Sci.* **5**, 1178450 (2023).
52. Daugman, J. G. & Downing, C. J. Demodulation, predictive coding, and spatial vision. *J. Opt. Soc. Am. A* **12**, 641–660 (1995).

53. Schofield, A. J. What does second-order vision see in an image? *Perception* **29**, 1071–1086 (2000).
54. Schmalfluss, J., Mehl, L. & Bruhn, A. Distracting downpour: adversarial weather attacks for motion estimation. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* (eds Agapito, L. et al.) 10106–10116 (IEEE, 2023).
55. Chubb, C. & Sperling, G. Two motion perception mechanisms revealed through distance-driven reversal of apparent motion. *Proc. Natl Acad. Sci. USA* **86**, 2985–2989 (1989).
56. Angelaki, D. E. & Hess, B. J. Self-motion-induced eye movements: effects on visual acuity and navigation. *Nat. Rev. Neurosci.* **6**, 966–976 (2005).
57. Fencsik, D. E., Klieger, S. B. & Horowitz, T. S. The role of location and motion information in the tracking and recovery of moving objects. *Percept. Psychophys.* **69**, 567–577 (2007).
58. Land, M. F. & Lee, D. N. Where we look when we steer. *Nature* **369**, 742–744 (1994).
59. Gershman, S. J., Tenenbaum, J. B. & Jäkel, F. Discovering hierarchical motion structure. *Vis. Res.* **126**, 232–241 (2016).
60. Bill, J., Gershman, S. J. & Drugowitsch, J. Visual motion perception as online hierarchical inference. *Nat. Commun.* **13**, 7403 (2022).
61. Jones, J. P., Stepnoski, A. & Palmer, L. A. The two-dimensional spectral structure of simple receptive fields in cat striate cortex. *J. Neurophysiol.* **58**, 1212–1232 (1987).
62. Jones, J. P. & Palmer, L. A. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiol.* **58**, 1233–1258 (1987).
63. Adelson, E. H. & Bergen, J. R. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A* **2**, 284–299 (1985).
64. Castet, E. & Zanker, J. Long-range interactions in the spatial integration of motion signals. *Spat. Vis.* **12**, 287–307 (1999).
65. Heeger, D. J. Modeling simple-cell direction selectivity with normalized, half-squared, linear operators. *J. Neurophysiol.* **70**, 1885–1898 (1993).
66. Carandini, M. & Heeger, D. J. Summation and division by neurons in primate visual cortex. *Science* **264**, 1333–1336 (1994).
67. Lu, Z. L., Lesmes, L. A. & Sperling, G. The mechanism of isoluminant chromatic motion perception. *Proc. Natl Acad. Sci. USA* **96**, 8289–8294 (1999).
68. Pack, C. C. & Born, R. T. Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain. *Nature* **409**, 1040–1042 (2001).
69. Gilaie-Dotan, S. Visual motion serves but is not under the purview of the dorsal pathway. *Neuropsychologia* **89**, 378–392 (2016).
70. Noest, A. & Van Den Berg, A. The role of early mechanisms in motion transparency and coherence. *Spat. Vis.* **7**, 125–147 (1993).
71. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems 30* (eds Guyon, I. et al.) 5998–6008 (NeurIPS, 2017).
72. Wang, X., Girshick, R., Gupta, A. & He, K. Non-local neural networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (eds Forsyth, D. et al.) 7794–7803 (IEEE, 2018).
73. Dosovitskiy, A. et al. An image is worth 16 × 16 words: transformers for image recognition at scale. In *International Conference on Learning Representation (ICLR)* (eds Mohamed, S. H. et al.) (2021).
74. Serre, T. Deep learning: the good, the bad, and the ugly. *Annu. Rev. Vis. Sci.* **5**, 399–426 (2019).
75. Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. Preprint at <https://arxiv.org/abs/1412.3555> (2014).
76. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)* (eds Ranzato, M. A. et al.) (2017).
77. Sun, D., Yang, X., Liu, M.-Y. & Kautz, J. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (eds Forsyth, D. et al.) 8934–8943 (IEEE, 2018).
78. Liu, L. et al. Learning by analogy: reliable supervision from transformations for unsupervised optical flow estimation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (eds Liu, C. et al.) 6489–6498 (IEEE, 2020).
79. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 834–848 (2017).
80. Wang, X., Girdhar, R., Yu, S. X. & Misra, I. Cut and learn for unsupervised object detection and instance segmentation. In *Proc. IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)* (eds Geiger, A. et al.) 3124–3134 (IEEE, 2023).
81. Huang, Z. et al. FlowFormer: a transformer architecture for optical flow. In *Proc. Computer Vision—ECCV 2022: 17th European Conference on Computer Vision* (eds Vedaldi, A. et al.) 668–685 (Springer, 2022).
82. Weiss, Y., Simoncelli, E. P. & Adelson, E. H. Motion illusions as optimal percepts. *Nat. Neurosci.* **5**, 598–604 (2002).
83. Sun, Z., Chen, Y., Yang, Y. & Nishida, S. Code of machine learning modeling for multi-order human visual motion processing. *Zenodo* <https://doi.org/10.5281/zenodo.14958959> (2025).
84. Achanta, R. et al. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**, 2274–2282 (2012).
85. Greff, K. et al. Kubric: a scalable dataset generator. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (eds Dana, K. et al.) 3749–3761 (IEEE, 2022).
86. Coumans, E. & Bai, Y. *Pybullet, A Python Module for Physics Simulation for Games, Robotics and Machine Learning* (2016); <https://pybullet.org>
87. *Blender—A 3D Modelling and Rendering Package* (Blender Online Community, 2021).
88. Zaal, G. et al. *Poly Haven: A Curated Public Asset Library for Visual Effects Artists and Game Designers* (Poly Haven, 2021).
89. *Unsplash Lite Dataset v1.3.0* (Unsplash, accessed 20 April 2025); <https://unsplash.com/data>
90. Yang, Y.-H., Sun, Z., Fukiage, T. & Nishida, S. HuPerFlow: a comprehensive benchmark for human vs. machine motion estimation comparison. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (eds Isola, P. et al.) 22799–22808 (CVF, 2025).
91. DeAngelis, G. C., Ohzawa, I. & Freeman, R. D. Receptive-field dynamics in the central visual pathways. *Trends Neurosci.* **18**, 451–458 (1995).

Acknowledgements

This work was supported in part by the Spring Fellowship (grant no. JPMJFS2123 to Z.S., Y.-J.C. and Y.L.) and in part by JSPS Grants-in-Aid for Scientific Research (KAKENHI) (grant nos. JP20H00603, JP20H05950, JP20H05957 and 24H00721 to S.N.). We thank Kubric⁸⁵ for providing their data generation pipeline.

Author contributions

Z.S. and S.N. conceived and designed the study. Z.S., Y.-J.C. and Y.-H.Y. performed data collection and preprocessing. Z.S. and Y.-J.C. developed the analysis pipeline and carried out statistical analyses. Y.L. generated segmentation results and visualizations. Z.S. drafted the paper. S.N. supervised the project, secured funding and critically revised the paper. All authors read and approved the final version.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s42256-025-01068-w>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42256-025-01068-w>.

Correspondence and requests for materials should be addressed to Shin'ya Nishida.

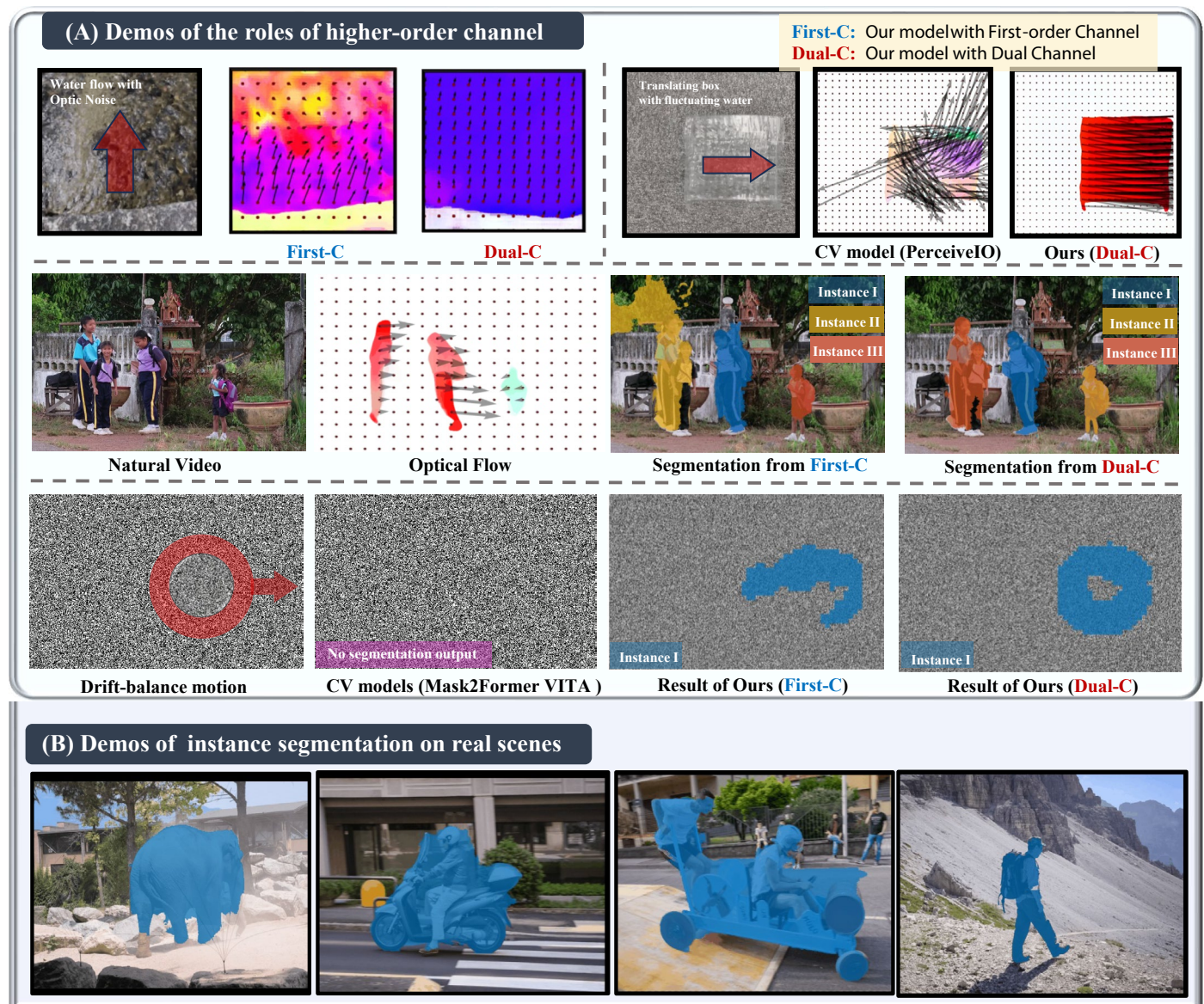
Peer review information *Nature Machine Intelligence* thanks Reuben Rideaux and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

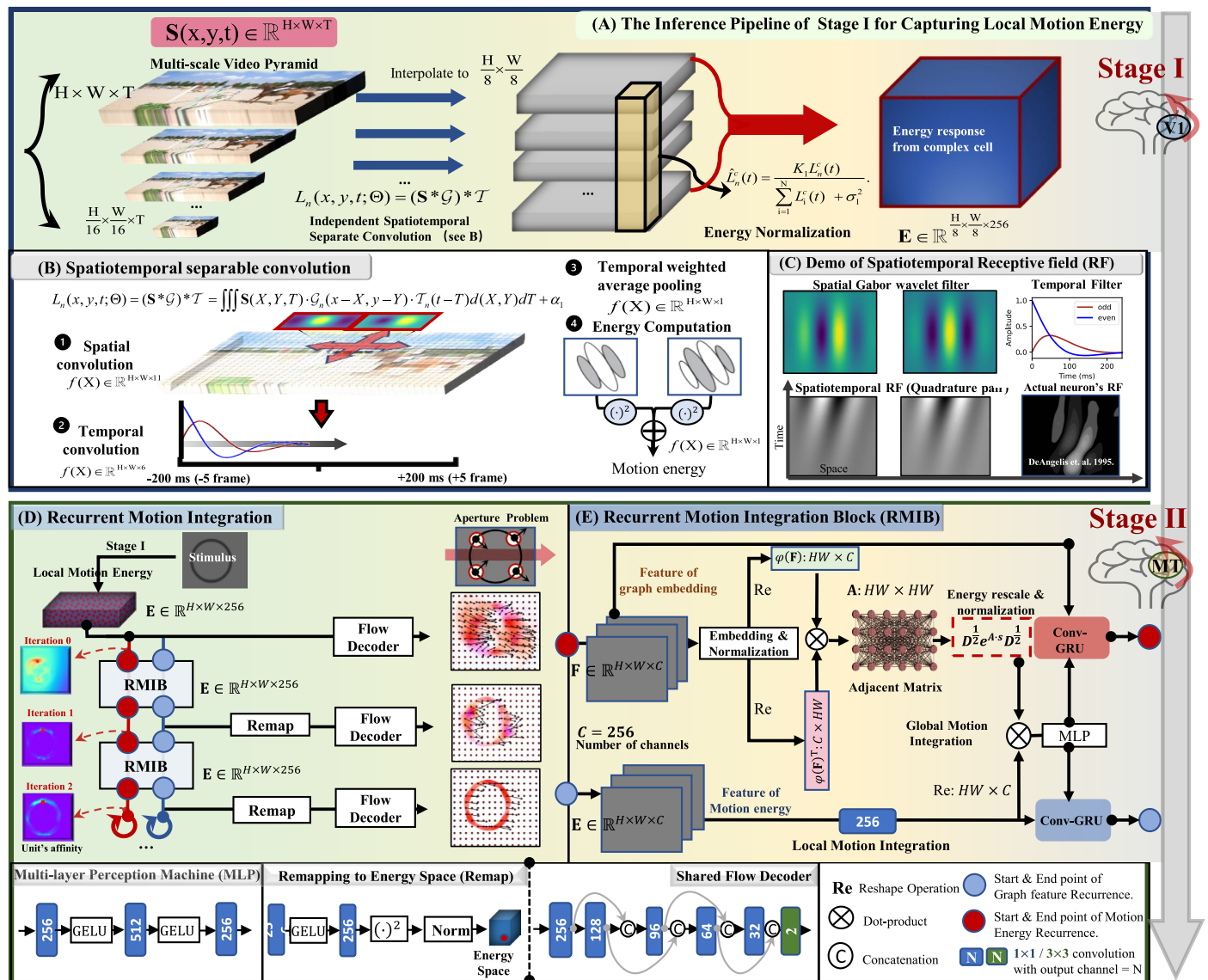
Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025



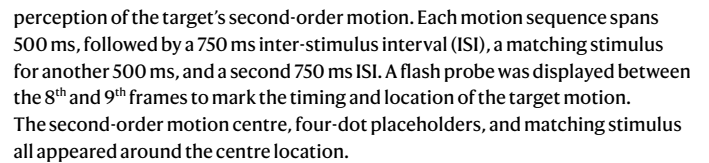
Extended Data Fig. 1 | Demonstration of the Roles Played by the Higher-Order Channel. Qualitative comparison of the first-order and dual-channel approaches. (A): First row: Motion estimation in noisy natural conditions, compared to a state-of-the-art CV method⁴⁷. The dual-channel model effectively suppresses water fluctuation noise, an example of second-order motion in natural scenes. Second row: Instance segmentation based on motion features,

where the dual-channel approach yields finer object segmentation. Third row: The higher-order channel improves the segmentation of drift-balanced motion, which remains undetected by SOTA CV segmentation methods⁵⁰. (B): Demonstration on motion segmentation in real scenes⁴⁸. Notably, all segmentation results are generated from the model's inherent graph structure, operating in a training-free, zero-shot manner.



Extended Data Fig. 2 | Modelling the motion perception system. Stage-I demonstrates the inference process for motion energy computation in the first-order channel. The higher-order motion channel follows a similar process, with an additional block of multilayer 3D convolutions and ReLU nonlinearity preceding the motion energy computation. This extension extracts higher-order nonlinear features. (a): The first stage employs a set of trainable motion energy units to capture local motion energy. (b): Motion energy computation, using

spatiotemporal separable filters, as a subcomponent of (b). (c): Illustration of spatiotemporal separable filters, including a quadrature pair of spatial filters and temporal filters. (d): The second stage uses a motion graph network with recurrent processing to simulate global motion integration and segregation, employing a flow decoder to visualize dense optical flow across iterations. (e): Global motion integration based on a motion graph and self-attention mechanism, as a subcomponent of (d).



Extended Data Table 1 | Model v.s. Human v.s. ground truth on KITTI 2015 Benchmark

Method	ρ_{uv}	ρ_{dir}	ρ_{spd}	RCI	v.s. Human				v.s. GT			
					r_{uv}	r_{spd}	r_{dir}	epe	r_{uv}	r_{spd}	r_{dir}	epe
Farneback [36]	0.06	-0.09	0.08	.039	0.79	0.29	0.42	5.56	0.92	0.68	0.86	2.34
FlowNet2.0 [6]	0.00	0.36	0.00	.025	0.83	0.32	0.61	5.27	0.98	0.92	0.81	1.10
RAFT [39]	-0.03	0.24	0.00	.012	0.84	0.34	0.57	5.19	1.00	0.99	0.89	0.49
AGFlow [40]	0.06	0.24	0.00	.050	0.77	0.22	0.52	5.26	0.90	0.63	0.70	2.15
FlowFormer [81]	-0.02	0.08	-0.03	.011	0.84	0.34	0.52	5.17	1.00	0.99	0.93	0.42
FFV1MT [37]	0.21	0.28	0.07	.071	0.80	0.27	0.47	5.37	0.88	0.64	0.51	3.39
VideoFlow [43]	0.01	0.06	0.02	.013	0.84	0.34	0.51	5.16	1.00	0.98	0.92	0.48
Peceiver IO [45]	0.04	-0.11	0.10	.023	0.84	0.36	0.49	5.21	0.99	0.94	0.96	0.95
Ours-first	0.27	0.36	0.20	.055	0.85	0.39	0.62	5.07	0.96	0.83	0.87	2.26

ρ : Partial correlation between human and model while controlling the effects of ground truth; r : Pearson correlation coefficient; epe : vector endpoint error; uv , dir , spd represent motion components in Cartesian space, direction, and speed, respectively; RCI is introduced from⁷ to represent the model's similarity to human perception (the larger, the more human-aligned). Human perceptual data are adopted from⁸⁸. 'Ours-first' represents the case in which we only use the first-order channel signal to generate the response.

Extended Data Table 2 | Model vs. Human on Second-order Motion

Method	Mod1		Mod2		Mod3		Mod4		Mod5		Mod6		Mod7		All	
	r_{uv}	EPE	r_{uv}	EPE	r_{uv}	EPE	r_{uv}	EPE	r_{uv}	EPE	r_{uv}	EPE	r_{uv}	EPE	r_{uv}	EPE
Trained with non-diffuse																
RAFT [39]	0.52	8.88	0.88	4.10	0.25	11.02	0.14	10.80	0.22	10.15	0.61	7.50	-0.14	10.89	0.35	9.05
MEGraph [16]	0.22	10.53	0.52	8.06	0.40	10.38	0.60	14.21	0.30	10.42	0.81	6.76	0.21	9.97	0.44	10.05
GMFlow [41]	0.07	10.80	0.09	9.11	0.07	10.45	0.11	10.82	0.16	10.34	0.13	10.05	-0.02	10.33	0.09	10.27
Videoflow [43]	0.54	8.88	0.51	7.62	0.57	10.05	0.19	10.77	0.56	10.73	0.74	5.76	-0.07	24.64	0.43	11.21
Ours-ND-dual	0.87	8.27	0.96	10.57	0.82	10.46	0.88	9.68	0.94	10.79	0.90	10.16	0.95	13.85	0.90	10.54
Trained without non-diffuse																
Farneback [36]	-0.04	11.14	-0.16	9.17	0.15	10.56	0.10	10.86	0.26	9.99	0.36	8.67	0.08	10.60	0.11	10.14
FFV1MT [37]	-0.26	11.26	-0.07	9.21	0.12	10.46	0.05	11.31	-0.02	10.83	0.01	11.06	-0.10	11.96	-0.04	10.87
RAFT [39]	0.31	10.55	0.39	8.11	0.22	10.85	0.14	10.72	0.16	10.30	0.44	8.06	0.02	10.33	0.24	9.84
MEGraph [16]	0.15	10.85	0.15	9.13	0.03	11.08	0.17	15.04	0.03	10.69	0.55	8.09	-0.17	11.56	0.13	10.92
GMFlow [41]	0.33	9.99	0.23	8.99	-0.19	10.49	0.11	10.78	0.06	10.40	0.04	10.48	-0.07	10.40	0.07	10.22
Videoflow [43]	0.33	9.98	0.25	8.60	0.19	12.31	0.38	10.81	0.18	10.50	0.43	8.80	0.16	10.31	0.27	10.19
Ours-D-dual	0.41	10.59	0.30	8.83	0.53	12.25	0.56	19.41	0.46	13.68	0.68	11.61	0.62	9.53	0.51	12.27
Ours Final																
Ours-first	-0.11	10.93	0.46	7.97	0.36	10.06	0.45	13.34	0.36	10.34	0.80	5.29	0.14	9.77	0.35	9.67
Ours-dual	0.91	3.95	0.93	3.34	0.90	3.81	0.93	3.14	0.95	2.97	0.95	2.71	0.96	2.59	0.93	3.22

For each dataset, we directly report the Pearson correlation (r_{uv}) and the vector endpoint error (EPE). We both show the method trained on diffuse and non-diffuse motion datasets. Mod1 to Mod7 are various second-order modulations, including random noise, Gaussian blur, water wave, Fourier phase shuffle, random pixel shuffle, swirl, and drift-balanced motion. The last two rows indicate our final jointly trained model, evaluated separately using a single first-order channel and a dual-channel approach. ‘Ours-ND’ and ‘Ours-D’ represent our model separately trained with non-diffuse and diffuse data.

Extended Data Table 3 | Model vs. Ground-truth on Second-order Motion

Method	Mod1		Mod2		Mod3		Mod4		Mod5		Mod6		Mod7		All	
	r_{uv}	EPE	r_{uv}	EPE	r_{uv}	EPE	r_{uv}	EPE	r_{uv}	EPE	r_{uv}	EPE	r_{uv}	EPE	r_{uv}	EPE
Trained with non-diffuse																
RAFT [39]	0.52	9.32	0.86	4.04	0.23	12.28	0.10	11.71	0.19	11.37	0.61	8.22	-0.15	12.22	0.34	9.88
MEGraph [16]	0.23	11.26	0.54	9.37	0.43	10.47	0.57	14.08	0.31	10.97	0.82	6.41	0.18	11.31	0.44	10.55
GMFlow [41]	0.05	11.58	0.09	11.58	0.06	11.65	0.11	11.65	0.16	11.53	0.12	11.47	-0.04	11.65	0.08	11.59
Videoflow [43]	0.55	9.21	0.49	9.30	0.56	10.50	0.20	11.59	0.56	11.17	0.74	5.54	-0.07	25.49	0.43	11.83
Ours-ND-dual	0.88	7.86	0.97	8.38	0.85	9.53	0.89	9.05	0.95	9.56	0.90	9.12	0.96	12.63	0.91	9.45
Trained without non-diffuse																
Farneback [36]	-0.03	11.87	-0.13	11.65	0.16	11.66	0.12	11.59	0.28	11.17	0.36	9.93	0.10	11.89	0.12	11.39
FFV1MT [37]	-0.26	12.01	-0.07	11.69	0.10	11.67	0.04	12.20	0.00	12.04	-0.02	12.45	-0.11	13.28	-0.05	12.19
RAFT [39]	0.32	11.07	0.36	10.50	0.19	12.11	0.12	11.64	0.15	11.50	0.45	9.20	0.03	11.65	0.23	11.09
MEGraph [16]	0.11	11.59	0.15	11.61	0.03	12.22	0.18	15.61	0.03	11.85	0.55	8.84	-0.15	12.88	0.13	12.08
GMFlow [41]	0.26	10.77	0.20	11.47	-0.15	11.69	0.09	11.61	0.05	11.60	0.06	11.89	-0.09	11.77	0.06	11.54
Videoflow [43]	0.32	10.70	0.26	10.92	0.18	13.44	0.33	11.63	0.17	11.63	0.42	9.98	0.13	11.64	0.26	11.42
Ours-D-dual	0.41	11.15	0.31	11.31	0.51	12.54	0.56	19.13	0.47	14.00	0.68	11.32	0.64	9.38	0.51	12.69
Ours Final																
Ours-first	-0.08	11.68	0.42	10.39	0.35	10.86	0.45	13.37	0.37	11.08	0.80	5.89	0.14	11.10	0.35	10.63
Ours-dual	0.93	4.06	0.93	3.37	0.91	4.17	0.94	3.11	0.95	2.61	0.95	2.92	0.97	2.58	0.94	3.26

For each dataset, we directly report the Pearson correlation (r_{uv}) and the vector endpoint error (EPE). We both show the method trained on diffuse and non-diffuse motion datasets. Mod1 to Mod7 are various second-order modulations, including random noise, Gaussian blur, water wave, Fourier phase shuffle, random pixel shuffle, swirl, and drift-balanced motion. The last two rows indicate our final jointly trained model, evaluated separately using a single first-order channel and a dual-channel approach. ‘Ours-ND’ and ‘Ours-D’ represent our model separately trained with non-diffuse and diffuse data.

Extended Data Table 4 | Model vs. Human vs. ground truth on First-order Motion

Method	Sintel						KITTI2015					
	to Human		to GT		ρ_{uv}	RCI	to Human		to GT		ρ_{uv}	RCI
	r_{uv}	EPE	r_{uv}	EPE			r_{uv}	EPE	r_{uv}	EPE		
Ours-first	0.91	0.93	0.90	0.88	0.50	0.67	0.85	5.07	0.96	2.26	0.27	.055
Ours-dual	0.88	1.12	0.90	0.93	0.29	.043	0.84	5.11	0.98	1.61	0.11	.040

We compare our first-order and dual-channel variants of our model on KITTI 2015³³ and Sintel³¹. For each dataset, we report the Pearson correlation (r_{uv}) and endpoint error (EPE) relative to both human judgments and ground truth. We also include the partial correlation (ρ_{uv}) between human and model while controlling for ground truth, along with the Response Consistency Index (RCI) from⁷. Notably, although adding the higher-order channel (Ours-dual) helps estimate the ground truth in some challenging scenes (KITTI), it reduces correlation with human data and decreases partial correlation, indicating that motion energy-based first-order motion processing is necessary and sufficient to capture human perceptual bias for these scenes. Unlike the first-order channel, the higher-order simply employs 3D convolution without human-aligned computational constraints and thus has more degrees of freedom, often introducing effects that deviate from human-like processing. This is also consistent with the fact that the Sintel slow subset and the validated KITTI regions generally lack non-diffuse reflections, making higher-order motion cues less relevant.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- | | | |
|-------------------------------------|-------------------------------------|--|
| ☐ | ☒ | The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☒ | ☐ | The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| ☐ | ☒ | A description of all covariates tested |
| ☒ | ☐ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☒ | ☐ | For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| ☒ | ☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ | Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	The data collection for the human psychophysical experiment was conducted using Python 3.9.12, along with the open-source libraries PsychoPy 2023.2, EasyDict 1.10, Pandas 2.0.0, and NumPy 1.23.5.
Data analysis	The data analysis and visualization was performed using Matlab 2023a and Python 3.9.12 with the following open-source libraries: NumPy 1.23.5, Pandas 2.0.0, Matplotlib 3.7.5, Seaborn 0.13.2, SciPy 1.7.3, and Pingouin 0.5.3. The code is available at: https://github.com/anonymized/multi-order-motion-model , also recorded at Zenodo: https://zenodo.org/records/14958959 .

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The project website is publicly available at: <https://anonymized.github.io/motion-model-website/>.

Human psychophysical data and the corresponding model responses are available at: <https://github.com/anoymized/multi-order-motion-model>, and are also archived on Zenodo at <https://zenodo.org/records/14958959>. All other relevant data supporting the findings of this study—including model predictions, human behavioral responses, and custom datasets (Drifting Grating, Non-textured 2D Motion, Diffuse Motion, Non-diffuse Motion, and Second-order Motion datasets)—are provided at the same repository. Two additional mini motion datasets featuring diffuse and non-diffuse objects are also made available to support quick verification of the effects on second-order motion perception. The public datasets used in this study are accessible from the following sources:

Kubric: <https://github.com/google-research/kubric>

KITTI: <https://www.cvlibs.net/datasets/kitti/>

MPI-Sintel: <http://sintel.is.tue.mpg.de/>

Sintel-slow: https://www.cvlibs.net/projects/slow_flow/

DAVIS: <https://davischallenge.org/>

Unsplash: <https://github.com/unsplash/datasets>

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender

Three participants (three male in sex) in the study were recruited, including two of the author and one naive participants, with an average age of 25.3 years (SD = 1.24). Participants self-report to be normal or corrected-to-normal vision. No other information related to other sex or gender-related information (including but not limited as gender, gender identity, etc) is collected.

Population characteristics

All the participants came from the Graduate School of Informatics Kyoto University, Japan. No other population characteristics were collected or recorded during the experiment.

Recruitment

The participants voluntarily joined the experiment and came from the Graduate School of Informatics, Kyoto University. The honorarium is given after quitting the experiment (including finishing or quitting the experiment). All the participants read and agreed with the consent form before the start of the experiment. All participants were told all their rights during the experiment, including the free-to-leave.

Ethics oversight

The experiment protocol, content, and consent form follow the Declaration of Helsinki and were screened and approved by the Ethics Committee of Kyoto University (approval no. KUIS-EAR-2020-003).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

This study is a quantitative experimental study. It involves collecting numerical data from controlled psychophysical experiments to evaluate participants' ability to estimate second-order motion vectors.

Research sample

This study used three male participants (average age: 25.3 years, SD = 1.24) with normal or corrected-to-normal vision. Two were authors of the study, and one was a naive participant from Kyoto University. All were graduate students recruited from the Graduate School of Informatics, Kyoto University, and were financially compensated. They were pre-trained with the task to ensure reliable data collection. Our choice is reasonable considering that low-level visual perception phenomena are widely regarded as highly conserved across healthy adults, regardless of sex or cultural background, and that the current purpose was to demonstrate well-established humans' high capability in perceiving second-order motion.

Sampling strategy

Our sampling procedure was based on convenience, as participants with normal or corrected-to-normal vision were recruited from the Graduate School of Informatics, Kyoto University. Although we did not pre-compute sample size, based on our past experiences with the measurement method, we designed our human experiment to measure a sufficiently large number of responses (1120 trials) for each participant to reliably compute the correlation with the ground truth. The results indicate that the responses were highly consistent within each participant, and across participants (mean correlation = 0.983, SD = 0.005, n = 3).

Data collection

Data were collected using a VIFWPixx /3D LCD monitor (VPixx Technologies) with a resolution of 1920x1080 pixels and a refresh rate of 30 Hz. Display luminance was linearly calibrated using an i1Pro chromometer. Participants used a mouse to provide responses during the experiment, which was programmed and run using the PsychoPy library. The experiments were conducted in a darkened room, and participants' heads were stabilized using a chinrest to ensure accurate viewing distance (70 cm). Only the participant and the researcher were present during the sessions. The researcher was not blind to the experimental conditions or hypotheses, as two of the participants were also authors of the study.

	However, the experimental protocol and stimulus presentation were fully randomized to reduce potential biases.
Timing	Data collection occurred on two separate dates. On October 16, 2023, experimental data from two authors were collected. Subsequently, data from the naive participants were collected on February 7, 2024. All sessions were completed within a single day for each participant.
Data exclusions	No data were excluded from the analyses. All collected data were included in the final analysis.
Non-participation	No participants dropped out or declined participation. All three participants who were recruited completed the experiment successfully.
Randomization	Participants were not allocated into experimental groups. The study followed a within-subject design where all participants completed the same set of trials.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems		Methods	
n/a	Involved in the study	n/a	Involved in the study
☒	☐ Antibodies	☒	☐ ChIP-seq
☒	☐ Eukaryotic cell lines	☒	☐ Flow cytometry
☒	☐ Palaeontology and archaeology	☒	☐ MRI-based neuroimaging
☒	☐ Animals and other organisms		
☒	☐ Clinical data		
☒	☐ Dual use research of concern		