MindSight: A Bio-Inspired Neural Architecture for Visual Restoration via Cortical Electrical Stimulation

Yongjie Zou[1], Haonan Niu[1], Bin Zhao[1], Guoliang Yi[1], Mengchuanzhi Yang[1], Jiawei Ju[1,2], Jiapeng Yin[1,2*], Chengyu T. Li[1*]

[1]Lingang Laboratory
[2]Shanghai Center for Brain Science and Brain-Inspired Technology
{zouyj, haonan.niu, zhaobin, yiguoliang, yangm, jiawei.ju, jiapeng.yin, tonylicy}@lglab.ac.cn

Abstract

Visual impairment is a common condition worldwide, and cortical electrical stimulation is one of the approaches to aid in visual restoration. However, existing methods suffer from limited precision, flexibility, and generalization in generating the desired visual perception. In this paper, we propose a novel deep learning-based algorithm for cortical electrical stimulation, named "MindSight," aimed at enhancing the clarity and accuracy of induced visual perceptions. Our framework introduces three key innovations: (1) A differentiable biophysical model simulating cortical state transitions under electrical stimulation, enabling end-to-end training; (2) A dual-path training architecture combining neural decoding fidelity with phosphene simulation constraints; (3) An attention-guided background gated network for input filtration and, a multi-channel activation constraint to ensure the effectiveness of electrical stimulation. We validated our approach through novel experiments with macaque monkeys, demonstrating superior performance in visual perception tasks. These results highlight the potential of our approach in assisting individuals with visual impairments.

Code — https://github.com/zyj9902/MindSight

Introduction

Visual impairment is a common condition worldwide (Stevens et al. 2013). The restoration of visual perception through cortical stimulation represents a transformative frontier in neuroprosthetics, offering hope for individuals with severe visual impairments caused by retinal degeneration or optic nerve damage. Traditional approaches, such as retinal implants (Dorn et al. 2013; Stingl et al. 2015), rely on intact neural pathways from the retina to the visual cortex, limiting their utility for patients with broader damage (Fernandez and Robles 2024). This outcome has motivated a shift toward approaches that bypass the eye entirely. In particular, visual cortex stimulation has gained interest (Lozano et al. 2020; Chen et al. 2020; Maghami et al. 2014; Roelfsema, Denys, and Klink 2018; Dobelle 2000) as it can potentially benefit a wider range of blindness causes (e.g. optic nerve damage, glaucoma) where retinal devices are ineffective. Recent advances with high-channel-count microelectrode arrays (96–1024 channels) demonstrate that intracortical stimulation can evoke shape-like percepts in both primates and humans (Chen et al. 2020; Fernández et al. 2021). Primary visual cortex (V1) contains an ordered map of the visual field and relatively large surface area, making it an attractive target for electrode implants (Beauchamp et al. 2020; Beauchamp and Yoshor 2020; Lewis et al. 2015; Tehovnik and Slocum 2013; Chen et al. 2020) that could evoke patterned visual percepts. Although these types of visual prostheses may differ significantly in terms of the entry point into the visual system, they share the same fundamental mechanism of action: through electrical stimulation of small groups of neurons, they evoke the perception of spatially localized flashes of light, called phosphenes (Brindley and Lewin 1968; Foroushani, Pack, and Sawan 2018). Despite advancements, research lacks adaptive and dynamic strategies, requiring manual determination of critical visual features and transformation of these features into suitable electrical stimulation protocols (Chen et al. 2020). Moreover, existing automated methods suffer from limited precision, flexibility, and generalization in generating the desired complex visual perception based on actual visual stimuli (Bosking, Beauchamp, and Yoshor 2017).

This work presents a novel bio-inspired algorithm, named "**MindSight**", for dynamically orchestrating electrode stimulation patterns in visual cortex. Our work is built on five key components: (1) A differentiable biophysical model simulating cortical state transitions under electrical stimulation, enabling end-to-end training; (2) A hierarchical training framework integrating neural decoding and phosphene simulation constraints, extending hybrid autoencoder (Granley, Relic, and Beyeler 2022); (3) Background gated network (BGN) for input filtration; (4) Multi-channel activation constraint to encourage a sufficiently large subset of electrode channels to fire concurrently, rendering the stimulation effective and consistent (Oswalt et al. 2021; Bosking et al. 2022); (5) Comprehensive validation through novel primate experiments demonstrating state-of-the-art performance.

Related Work

In biological vision, two frequently explored directions are: (1) visual encoding mechanisms—how visual stimuli elicit neural responses; and (2) visual decoding algorithms—how

*Corresponding authors.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

to reconstruct perceived visual scenes from neural signals. Extensive research has been conducted on the neural encoding mechanisms and response characteristics to visual stimuli in rodents (Turishcheva et al. 2024b,a; Li et al. 2023; Xu et al. 2023; Sinz et al. 2018) and primates (Cadena et al. 2019; Hatanaka et al. 2022; Farzmahdi, Kohn, and Coen-Cagli 2025; Shipp 2024), leading to significant advances and a clearer understanding of these processes. Similarly, substantial progress has also been made in the area of visual decoding. Pixel-level reconstructions of images from brain signals demonstrate how encoding and decoding approaches inform each other (Zhang et al. 2020, 2022). Meanwhile, other studies focus on decoding semantic information (Chen et al. 2023; Ozcelik et al. 2022; Gaziv et al. 2022).

Given that the aforementioned studies primarily focus on the relationship between visual input and neural activity, either predicting natural neural responses based on given visual stimuli and brain states, or conversely decoding perceived visual input from neural activity. However, they differ substantially from the goal of visual prosthetics: determining how to appropriately intervene in the visual system to evoke desired percepts. Nonetheless, the mechanisms of visual encoding and decoding still provide important biological insights into how visual perception might be restored.

Early visual prosthetics focused on retinal stimulation, exemplified by systems like Argus II, which translates camera-captured images into electrical signals for retinal neurons (Luo and Da Cruz 2016). However, such devices require intact retinal circuitry, rendering them ineffective for patients with optic nerve damage or advanced degenerative diseases. Cortical visual prosthetics bypass these limitations by directly stimulating visual cortex neurons but initially yielded only coarse visual perceptions (Brindley and Lewin 1968; Dobelle, Mladejovsky, and Girvin 1974). There is a need to explore more complex pattern presentation (Dobelle and Mladejovsky 1974; Bosking, Beauchamp, and Yoshor 2017). Recent cortical prosthetic approaches advanced visual perception by either sequentially activating electrodes to mimic natural motion-sensitive processing (Beauchamp et al. 2020), enabling accurate letter recognition in humans, or employing high-channel-count simultaneous stimulation to evoke shape perception in monkeys (Chen et al. 2020). Meanwhile, some deep learning techniques have also been introduced into this field. A deep autoencoder-based architecture that includes a highly adjustable prosthetic vision simulation module attempts to automatically find a task-specific stimulation protocol (van Steveninck et al. 2022). Hybrid neural autoencoder (Granley, Relic, and Beyeler 2022) combines biological constraints with deep learning to generate stimulation that preserve topological relationships in V1. Another ML-based paradigm is to incorporate models of the visual system's response into the stimulus optimization process (Grani et al. 2022).

However, these methods either remain at the theoretical simulation stage without sufficient animal experiments to validate their effectiveness, or the visual percepts elicited by the methods are imprecise, inflexible, and unclear. To address these issues, this work proposes MindSight.

Methodology

Overview

As shown in Figure 1, MindSight consists of three major stages. The first stage is the MUA-Image Decoder (MID), which decodes neural activity in the form of multi-unit activity (MUA) into corresponding visual images. The second stage is the Electrical Stimulation Generator (ESG), which optimizes the stimulation parameters to generate the visual perceptions as close as possible to the desired perceptions. This stage involves training through two paths of losses. The third stage, the Background Gated Network (BGN), filters out irrelevant visual stimuli, ensuring that electrical stimulation is applied only to stimuli that contain essential visual information. The biologically inspired designs of the individual modules, along with the rationale behind each, will be detailed in the corresponding subsections that follow. To provide an overview of how the algorithm operates as a whole, the high-level algorithmic workflow of the entire framework is presented in **Algorithm 1** in the **Appendix**.

Stage A: MUA-Image Decoder (MID)

In previous work, attempts were made to decode pixel-level images from retinal ganglion cells spikes (Zhang et al. 2020) and two-photon calcium signals of the Macaque Visual Cortex (Zhang et al. 2022). However, these models do not incorporate biological visual mechanisms. To reflect the selective tuning of visual cortical neurons to different visual attributes, the modulation of receptive fields, and the feedback and integration across visual cortical areas, MUA-Image Decoder (see the upper part of Figure 1) introduces three key advancements over previous neural decoders: (1) Spatiotemporal fusion of multi-scale MUA dynamics, (2) Hierarchical attention-guided feature refinement, and (3) Biologically-plausible feature recombination through adaptive skip connections. Our architecture incorporates the Convolutional Block Attention Module (CBAM) (Woo et al. 2018) for cross-dimensional feature refinement. As shown in Figure 1, the overall calculation process of MID is as follows:

$$\hat{I}_i = \mathcal{D}_\psi(\mathcal{E}_\phi(\mathcal{P}_\theta(\text{MUA}_i))) \tag{1}$$

Here, $\hat{I}_i$ denotes the image decoded from MUA (1280 channels × 50 ms), $\mathcal{P}_\theta$ represents the projection network, and $\mathcal{E}_\phi$ and $\mathcal{D}_\psi$ denote the encoder and decoder of the Image-Image autoencoder, respectively. Refer to **Appendix A** for details.

Stage B: Electrical Stimulation Generator (ESG)

The ESG (see the lower part of Figure 1) is responsible for producing voltage amplitudes on each electrode channel, given a real input image. The ESG parameters would be optimized via two paths (referred to as *Path 1* and *Path 2*) and one channel-activation constraint term, each contributing to the overall loss function:

1. **Path 1: MID Decoding Constraint.** The purpose of this constrained pathway is to ensure that the neural state changes and responses induced by cortical electrical stimulation can be decoded—via the MID trained in stage A—into the intended visual input. This provides

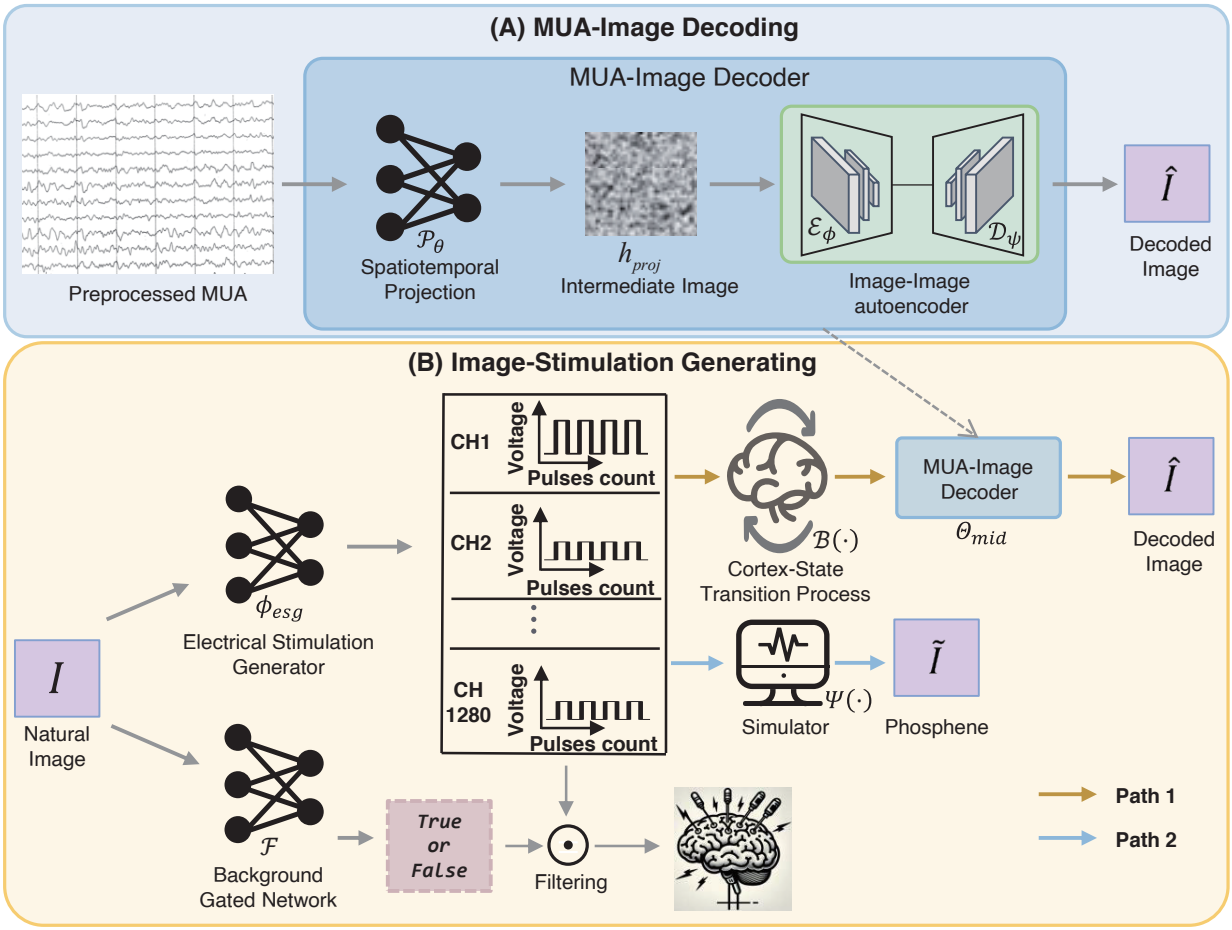

Figure 1: **MindSight.** Stage A (upper part): In MUA-Image Decoding, MUA-Image Decoder is trained to reconstruct visual stimuli from the multi-unit activity (MUA) recorded in the primary visual cortex, and the well-trained MUA-Image Decoder will be used in stage B. Stage B (lower part): In Image-Stimulation Generating, electrical stimulation generator (ESG) is responsible for producing voltage amplitudes on each electrode channel, given a real input image, and its parameters are trained via two paths. Background gated network (BGN) is trained to filter the external visual inputs, ensuring that only images containing key information will trigger cortical electrical stimulation.

a theoretical guarantee of stimulation effectiveness from the perspective of biological vision systems. Otherwise, how can we ensure that the visual percepts generated by electrical stimulation are meaningfully related to the desired visual inputs?

2. **Path 2: Phosphene Simulation Constraint.** The inclusion of this constrained pathway is also motivated by the need to satisfy established biological priors—namely, the orderly mapping between the visual cortex and phosphene perception. Without this, even at the coarse level of phosphene perception, the correctness of electrical stimulation cannot be ensured.
3. **Channel Activation Constraint.** To ensure that sufficiently many channels within the same electrode array are activated in unison. Without this, stimulating too few electrode channels in a local region may render the stimulation ineffective—that is, no percept may be elicited in the corresponding area of the visual field.

Architecture of ESG To better enable the ESG to learn the relationship between the target visual percept and the corresponding electrical stimulation parameters, it is essential to extract features from different regions of the visual target space and across multiple levels of granularity. At the same time, the design must account for the deployment and computational cost of ESG.

Depthwise-separable convolutions are employed to reduce the siz of ESG, aiming to minimize power consumption, as ESG will be integrated into the chip within the visual prosthesis system. Hierarchical processing with depthwise-separable convolutions and attention:

$$e_{out} = \text{AvgPool}(\text{Stack}_{4\times}[\text{ResBlock}](I)) \tag{2}$$

Use $\mathcal{R}$ to explicitly denote each ResBlock, as shown below:

$$\mathcal{R}(x) = \text{ReLU}\Big(\underbrace{\mathcal{C}_{\text{DS}}^{(2)}(\text{CBAM}(\text{ReLU}(\mathcal{C}_{\text{DS}}^{(1)}(x))))}_{\text{conv path}} + \underbrace{\mathcal{C}_{\text{proj}}(x)}_{\text{shortcut}}\Big) \tag{3}$$

where $\mathcal{C}_{\text{DS}}^{(i)}$ denotes depthwise-separable conv with BN.

After adaptive global pooling, we apply two fully connected layers (with LayerNorm and GELU) to predict stimulation parameters:

$$V = \underbrace{W_2\Big(\text{GELU}\big(\text{LayerNorm}(W_1 e_{\text{out}})\big)\Big)}_{\in \mathbb{R}^{K\times T_{\text{pulse}}}} \tag{4}$$

Here W_1 and W_2 are linear transformations mapping and V represents the voltage amplitudes for the K electrode channels. T_{pulse} denotes the number of pulse for each channel.

Path 1: MID Decoding Constraint As shown in Figure 1, in path 1, we introduce the cortex-state transition process and the previously trained MID to guide ESG: 1. A real input image I_i is fed into the ESG, denoted by $\phi_{esg}(\cdot)$, yielding a set of voltage amplitudes across all electrode channels. 2. A cortex-state transition process model simulates how these voltage amplitudes (converted to currents) affect cortical neural states over time, producing a simulated neural response. 3. The simulated neural response is passed through the fixed-parameter MID, which reconstructs an image $\hat{I}_i$.

Let Θ_{mid} be the fixed parameters of MID. Let $\mathcal{B}(\cdot)$ denote the *cortex-state transition process*, modeling the spatiotemporal changes in mua triggered by electrical stimulation. Path 1 can be summarized as:

$$\hat{I}_i = \Theta_{mid}\big(\mathcal{B}(\phi_{esg}(I_i))\big) \tag{5}$$

Cortex-State Transition Process Suppose each of the K electrodes is located at spatial coordinates (x_k, y_k, z_k), and let $\phi_{esg}^k(I_i)$ be the generated voltage amplitude on electrode k. If R_k is the impedance coefficient of channel k, then its injected current is approximately

$$I_k = \frac{\phi_{esg}^k(I_i)}{R_k} \tag{6}$$

We adopt Spatiotemporal Gaussian distribution to describe how stimulation at (x_k, y_k, z_k) and current I_k diffuses through cortical tissue over time. In the x, y, z directions, we assume that the diffusion rates α are the same, i.e., $\sigma_x(t) = \sigma_y(t) = \sigma_z(t) = \sqrt{\alpha t + \sigma_0^2}$. Specifically, after calculation and simplification, the electrical stimulation for channel k ultimately generated the following distribution:

$$\mathcal{G}_k(\mathbf{p}, t) = \frac{I_k}{(2\pi(\sigma_0^2 + \alpha t))^{\frac{3}{2}}} \exp\Big(-\frac{\|\mathbf{p} - (x_k, y_k, z_k)\|^2}{2\sigma_0^2 + 2\alpha t}\Big) \tag{7}$$

where $\mathbf{p}$ denotes a spatial coordinate, t represents time, σ_0 is the initial standard deviation, and α is the diffusion rate. Summing contributions from all K electrodes yields

$$\mathbf{S}(\mathbf{p}, t) = \sum_{k=1}^{K} \mathcal{G}_k\big(\mathbf{p}, t; I_k, (x_k, y_k, z_k)\big) \tag{8}$$

thus forming a spatiotemporal distribution of the electric field. By sampling $\mathbf{S}(\mathbf{p}, t)$ at the corresponding recording channels over a certain time window, we obtain the simulated neural response, which is then fed into the MID. Note that T_{pulse} is a critical hyperparameter and Equations (7) and (8) only describes the case where T_{pulse} is 1. If T_{pulse} is greater than 1, the complex effects introduced by different pulses must also be considered. In our experiments, we set T_{pulse} to 1. If you find that a small number of pulses cannot activate the channel, then T_{pulse} needs to be increased.

Loss Function for Path 1 The goal of path 1 is to ensure that the images reconstructed from simulated neural responses are highly consistent with the real input images, thereby ensuring that the electrical stimulation parameters generated by ESG can induce the desired visual perception. Using the mean squared error at the pixel level, we define

$$\mathcal{L}_{\text{Path1}} = \mathbb{E}[\|I - \hat{I}\|_2^2] \tag{9}$$

Path 2: Phosphene Simulation Constraint In path 2, the same generator $\phi_{esg}(\cdot)$ is used, but the forward pass relies on a *phosphene simulator* (van der Grinten et al. 2024) (see the lower part of Figure 1), denoted by $\Psi(\cdot)$, and $\Psi(\cdot)$ is not trainable. The pipeline is: 1. A real image I_i is given to the generator $\phi_{esg}(\cdot)$, producing voltage amplitudes (converted to currents). 2. Generated voltage amplitudes $\phi_{esg}^k(I_i)$, impedance coefficient R_k and the center positions of the receptive fields $(\tilde{x}_k, \tilde{y}_k)$ for all channels are fed into the phosphene simulator $\Psi(\cdot)$, which outputs a simulated phosphene image $\tilde{I}_i$. $(\tilde{x}_k, \tilde{y}_k)$ can be obtained through Receptive Field Mapping (see Section **Receptive Field Mapping** for details). Mathematically, we have

$$\tilde{I}_i = \Psi\big(\phi_{esg}^k(I_i), R_k, (\tilde{x}_k, \tilde{y}_k)\big) \tag{10}$$

For each real input I_i, we generate the corresponding binarized target $\dot{I}_i$ (see **Appendix D: Method for Constructing Binarized Targets** for more details). The path 2 loss is then:

$$\mathcal{L}_{\text{Path2}} = \mathbb{E}[\|\dot{I} - \tilde{I}\|_2^2] \tag{11}$$

which penalizes discrepancies between the simulated and target phosphene images to adhere to the orderly mapping between the visual cortex and phosphene perception.

Channel Activation Constraint Our extensive saccade experiments (Tehovnik, Slocum, and Schiller 2003; Tehovnik et al. 2005) on monkeys have shown that, stimulating too few electrode sites within a localized area may render the stimulation ineffective, i.e., no perception was induced in the corresponding visual field region (see Section **Electrical-Stimulation-Induced Saccade** for details). To encourage a sufficiently large subset of electrode channels to fire concurrently, we introduce an additional penalty:

$$\mathcal{L}_{\text{con}} = \gamma \mathbb{E}\left[\sum_k \exp\left(-\frac{\phi_{esg}^k(I)}{\epsilon R_k}\right)\right] \tag{12}$$

where γ and ϵ are the penalty coefficient and decay coefficient, respectively. This penalty term constrains the number of effective channels that are not stimulated to be as few as possible. In particular, when the physical locations of different channels are closer, the probability of them being stimulated simultaneously becomes higher (rather than stimulating only a subset of these channels, even though stimulating only this subset could already make $\mathcal{L}_{\text{Path2}}$ very small).

Overall Loss for the ESG By combining Path 1, Path 2, and the channel activation constraint, the parameters of ESG are trained by minimizing the total objective using Adam:

$$\mathcal{L}_{\text{gen}} = \lambda_1 \mathcal{L}_{\text{Path1}} + \lambda_2 \mathcal{L}_{\text{Path2}} + \mathcal{L}_{\text{con}}, \tag{13}$$

where λ_1 and λ_2 control relative importance of two paths.

Stage C: Background Gated Network (BGN)

Given that multiple studies (Cogan 2008; Aungaroon et al. 2017; Larkin et al. 2022; Vatsyayan and Dayeh 2022; Vatsyayan et al. 2021) have shown excessive cortical electrical stimulation can increase the risk of neuronal necrosis, tissue damage, electrode failure, and even trigger after-discharges or seizures, we further train a BGN (see the lower part of Figure 1) to filter visual inputs. This ensures that only images containing key information trigger cortical stimulation.

BGN consists of a ResNet backbone modified with CBAM layers to dynamically refine feature maps. Given I_i, BGN indicates whether the image contains key information:

$$p_i = \text{Sigmoid}\left(\text{FC}\left(\text{CBAM-ResNet}(I_i)\right)\right) \tag{14}$$

Experiment

Datasets

In our study, we used two self-collected datasets: **Monkey Neural-Image Viewing Dataset (NIVD)** and **Monkey Vehicle Obstacle Avoidance for Food — Neuro-Visual Dataset (MVOAF-NVD)**. Refer to **Appendix C** for more description of these two datasets. For details on how NIVD and MVOAF-NVD were collected, refer to **Appendix C.1**, **Appendix C.2**, and Supplementary video 1.

Implementation Details

For all hyperparameter choices and implementation details related to data preprocessing and the training of MindSight, please refer to **Appendix B**.

Receptive Field Mapping

We used a fast RF mapping method similar with (Fiorani et al. 2014) to get the receptive fields of neurons. Refer to **Appendix E** for more details.

Electrical-Stimulation-Induced Saccade

We followed previously reported methods (Chen et al. 2020) to conduct electrical-stimulation-induced saccade experiments on our monkeys. We found that at least 16 channels per electrode are required to elicit effective phosphene perception, which motivated the design of Channel Activation Constraint. For more saccade experiments details, refer to **Appendix F** and the corresponding Table S1.

Behavioral Validation Experiments in Monkeys

To validate the efficacy of MindSight, we designed a delayed match-to-sample (DMS) task and a driving-based obstacle-avoidance foraging (DOAF) task. Refer to **Appendix G** and Figure S1 for more description of these two tasks.

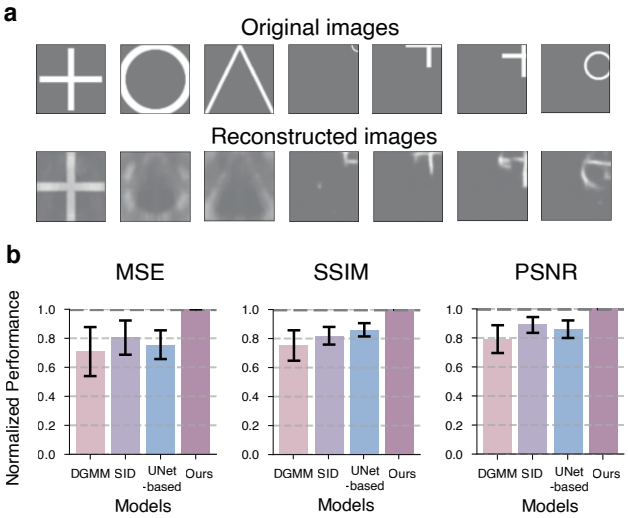

Figure 2: **Decoding Performance Comparisons on NIVD Test Set. a.** Some examples of decoded images. **b.** We evaluated four models (ours, DGMM, SID, and UNet-based) on the NIVD test set using three evaluation metrics: MSE, SSIM, and PSNR. Error bars represent standard deviation of the mean normalized performance.

Results

For MindSight, it is mainly necessary to validate the results of the following modules it contains: 1. Decoding performance of MID, as it indirectly affects the performance of ESG; 2. Phosphenes generated by ESG; 3. Performance of ESG in DMS; 4. Performance of ESG and BGN in DOAF.

Performance of MID

We evaluated the performance of MID on the test set of the **NIVD** (4,037 cropped images with corresponding MUAs). Figure 2a demonstrates MID's decoding results on some examples. As shown in Figure 2b, MID was compared with several state-of-the-art pixel-level decoding models—SID (Zhang et al. 2020), UNet-based model (Zhang et al. 2022), and DGMM (Du et al. 2018)—using evaluation metrics including mean square error (MSE) that describes the absolute difference of every pixel, the structural similarity index measure (SSIM) that captures the details or image distortion (Wang et al. 2004), and the peak-signal-to-noise-ratio (PSNR) that characterizes the global quality. For clarity, we utilized normalized model performance. Specifically, each model's score for every test sample was normalized against MID's scores, where the ratio for MID is defined as 1. Values above 1 indicate better performance than MID, while values below 1 indicate inferior performance. For the other three models, we calculated the mean and variance of their normalized performance across the entire test set. Compared to the other three models, MID from MindSight exhibited superior performance across all three metrics.

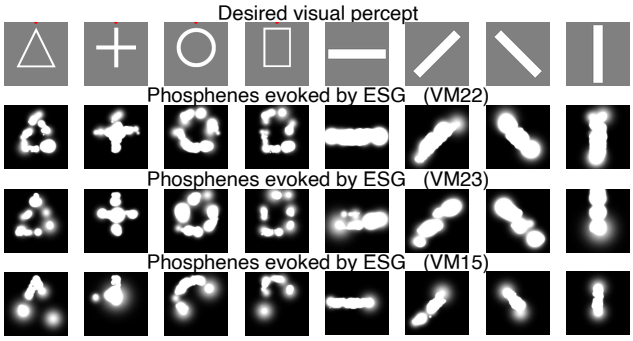

Figure 3: **Phosphenes Evoked by ESG in DMS.** The first row shows eight possible visual percepts in DMS. The second, third, and fourth rows show the phosphenes that ESG would theoretically evoke for VM22, VM23, and VM15, respectively, when targeting each desired visual percept.

Visual percepts	VM22		VM23		VM15	
	IoU	Dice	IoU	Dice	IoU	Dice
Triangle	0.60	0.75	0.57	0.73	0.43	0.60
Cross	0.64	0.78	0.68	0.81	0.48	0.65
Circle	0.61	0.76	0.57	0.73	0.48	0.64
Rectangle	0.52	0.68	0.46	0.63	0.36	0.53
bar (0°)	0.76	0.86	0.64	0.78	0.61	0.76
bar (45°)	0.69	0.82	0.73	0.84	0.58	0.74
bar (135°)	0.67	0.80	0.70	0.82	0.47	0.64
bar (90°)	0.67	0.80	0.60	0.75	0.49	0.66

Table 1: **Quantitative Results for Figure 3.** For each monkey, we compute the similarity between each desired visual percept and the corresponding phosphene evoked by ESG. Since the focus here is primarily on shape and contour, the original visual percept images are first preprocessed (e.g., binarization) before computing the Intersection over Union (IoU) and Dice Similarity Coefficient with the phosphenes.

Phosphenes Evoked by ESG

We also evaluated the simulated phosphenes evoked by ESG in both DMS and DOAF. As shown in Figure 3, we assessed the theoretical phosphenes that would be induced by the stimulation parameters generated by ESG when aiming to evoke different desired visual percepts in DMS for VM15, VM22, and VM23. The phosphenes evoked by ESG vary across monkeys (Table 1 provides a quantitative summary for each monkey), due to differences in electrode implantation sites, which in turn lead to variations in receptive fields (see **Appendix E** for more details). These differences in electrode and receptive field distributions influence the training of MindSight, as described in Equations (7), (8), and (10). The results for VM15 are significantly worse than those for VM22 and VM23, primarily due to the suboptimal distribution of electrode implantation sites. The phosphenes evoked by ESG in DOAF will be detailed in Section **Performance of ESG and BGN in DOAF**.

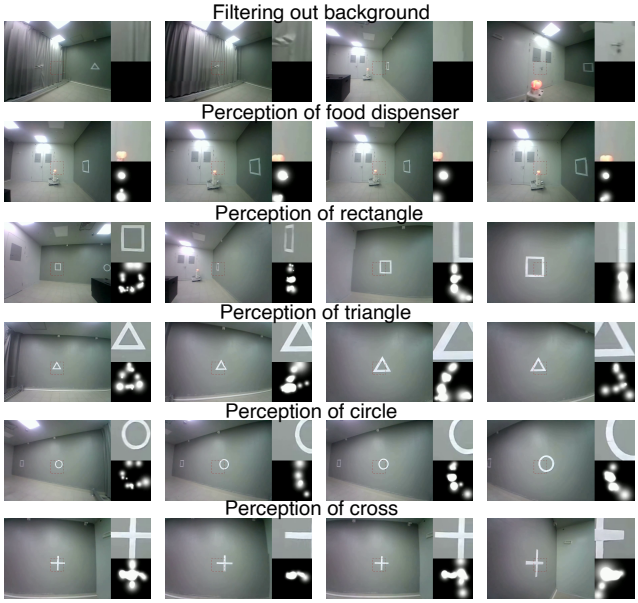

Figure 4: **Examples of Phosphenes Evoked by ESG in DOAF.** The first row shows the results of background filtering, where irrelevant information such as door handles, curtains, and walls does not trigger electrical stimulation. Rows two through six illustrate the stimulation evoked for the food dispenser and various markers at different angles and distances as the cart moves.

Performance of ESG in DMS

The ESG trained on NIVD was first validated in DMS (as outlined in Section **Behavioral Validation Experiments in Monkeys**). We evaluated model efficacy by measuring the monkey's behavioral success rate in DMS, where the choice relies entirely on electrical stimulation, with no visual cues provided. Figure S1 displays four-class DMS, with the two-class DMS differing only in the final selection step between two shapes instead of four shapes. Figure 5a shows the dynamic performance of behavioral success rates in a two-class DMS during one session. With electrical stimulation from the ESG, the monkey maintained high accuracy (red region). However, when the stimulation was removed, the success rate immediately dropped to chance level (green region). As shown in Figure 5b, the results of MindSight's ESG in two-class DMS were compared with those from another Chen's study (Chen et al. 2020). Both experiments followed nearly identical protocols, except that Chen used some different shapes, such as "T" and "L". We validated our method using two monkeys (monkeys 22 and 23), while Chen used monkeys A and L, with results statistically aggregated over 10 sessions per monkey. The comparison reveals that our method achieves higher stability. Moreover, compared to (Chen et al. 2020), MindSight also offers greater flexibility, as it can directly adapt to arbitrary inputs without the need for additional manual intervention. Such robustness and flexibility are crucial for the future application of visual prosthetics. We additionally demonstrated the mon-

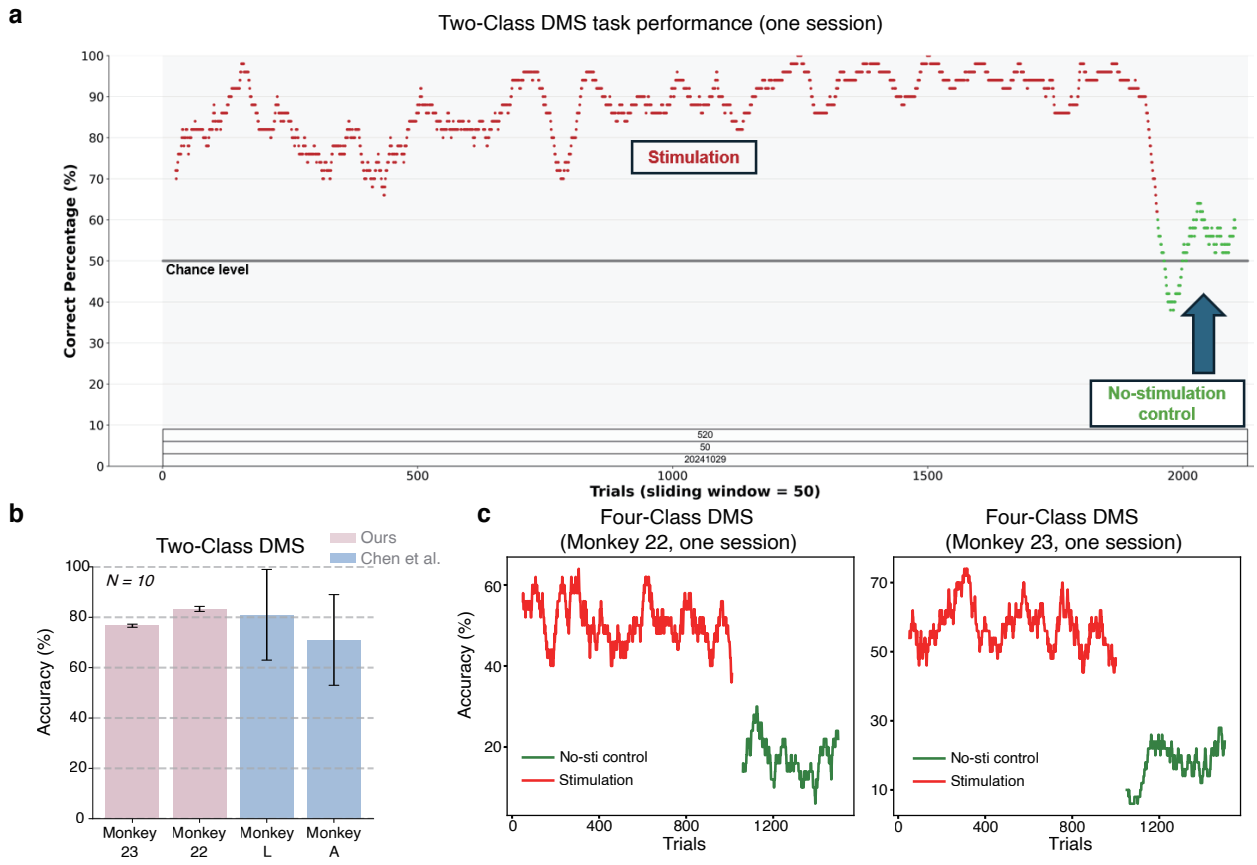

Figure 5: **Validation on the DMS Task. a.** Dynamic performance of MindSight in two-class DMS task during one session. **b.** Performance comparisons in two-class DMS task. Error bars represent the SEM (Standard Error of the Mean). N, number of sessions for each monkey and model. **c.** Dynamic performance of MindSight in four-class DMS task during one session.

keys' performance in four-class DMS (which has not been explored in (Chen et al. 2020)). Figure 5c illustrates the dynamic performance of behavioral success rates in one session for monkeys 22 and 23, respectively. Similarly, a control group without electrical stimulation was set up, confirming that MindSight's correct rate far exceeded the chance level (25%). Supplementary videos 2 and 3 demonstrate examples of MindSight applied to 2-class (Monkey 22) and 4-class (Monkey 23) DMS, respectively, showing the monkeys performing continuous correct trials. See **Appendix H** for details of Supplementary Videos 2–3.

Performance of ESG and BGN in DOAF

The ESG (trained on MVOAF-NVD) was validated in DOAF (as outlined in Section **Behavioral Validation Experiments in Monkeys**). Because a full session is lengthy, Supplementary Video 4 presents only segments highlighting MindSight's performance at key moments (e.g., when the monkey observes markers on walls, food dispensers, or irrelevant background). In the video, left panel shows head-mounted camera footage (on a head-fixed monkey), with the red dashed box representing a 64×64 region corresponding in both size and position to the receptive field covered by implanted electrodes. Upper-right display shows the content within this red dashed box, which serves as the input to ESG/BGN. Lower-right panel simulates the visual perception evoked by the electrical stimulation commands generated by ESG (Equation (10)), combined with BGN outputs. For easier viewing, Figure 4 presents some frames from Supplementary Video 4 as illustrative examples. BGN-triggered stimulation activates only during key cues (e.g., markers, obstacles, and food dispensers), filtering out irrelevant information to avoid the risks associated with excessive electrical stimulation. Notably, despite variations in the shape, angle, and position of markers and food dispensers as the cart moves, MindSight consistently evoked effective visual perceptions.

Conclusion

We present MindSight, a framework for visual restoration via cortical electrical stimulation. By integrating differentiable biophysical modeling with dual-path training, MindSight outperforms other approaches and shows good performance with comprehensive validation through novel primate experiments. For additional discussion, see **Appendix I**.

Acknowledgments

This work was supported by National Science and Technology Innovation 2030 Major program (No. 2021ZD0203601), National Natural Science Foundation of China (No. 32221003, No. 32221003), Shanghai Municipal Science and Technology Major Project (No. 2018SHZDZX05, No. 2021SHZDZX), National Key R&D Program Key Scientific Issues of Transformational Technology (No. 2019YFA0709504), Shanghai Pilot Program for Basic Research-Chinese Academy of Science, Shanghai Branch (No. JCYJ-SHFY2022-010), Lingang Laboratory (No. LG202105-01-01, No. LG202105-01-11, No. LG-GG-202402-06, No. LGL-5925), Shanghai Sailing Program (NO. 22YF1460900, No. 23YF1461000), Shanghai Postdoctoral Excellence Program (No. 2023021), and Shanghai Yang Fan Foundation (No. 24YF2730700).

References

Aungaroon, G.; Vera, A. Z.; Horn, P. S.; Byars, A. W.; Greiner, H. M.; Tenney, J. R.; Arthur, T. M.; Crone, N. E.; Holland, K. D.; Mangano, F. T.; et al. 2017. After-discharges and seizures during pediatric extra-operative electrical cortical stimulation functional brain mapping: incidence, thresholds, and determinants. *Clinical Neurophysiology*, 128(10): 2078–2086.

Beauchamp, M. S.; Oswalt, D.; Sun, P.; Foster, B. L.; Magnotti, J. F.; Niketeghad, S.; Pouratian, N.; Bosking, W. H.; and Yoshor, D. 2020. Dynamic stimulation of visual cortex produces form vision in sighted and blind humans. *Cell*, 181(4): 774–783.

Beauchamp, M. S.; and Yoshor, D. 2020. Stimulating the brain to restore vision. *Science*, 370(6521): 1168–1169.

Bosking, W. H.; Beauchamp, M. S.; and Yoshor, D. 2017. Electrical stimulation of visual cortex: relevance for the development of visual cortical prosthetics. *Annual review of vision science*, 3(1): 141–166.

Bosking, W. H.; Oswalt, D. N.; Foster, B. L.; Sun, P.; Beauchamp, M. S.; and Yoshor, D. 2022. Percepts evoked by multi-electrode stimulation of human visual cortex. *Brain stimulation*, 15(5): 1163–1177.

Brindley, G. S.; and Lewin, W. S. 1968. The sensations produced by electrical stimulation of the visual cortex. *The Journal of physiology*, 196(2): 479–493.

Cadena, S. A.; Denfield, G. H.; Walker, E. Y.; Gatys, L. A.; Tolias, A. S.; Bethge, M.; and Ecker, A. S. 2019. Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS computational biology*, 15(4): e1006897.

Chen, X.; Wang, F.; Fernandez, E.; and Roelfsema, P. R. 2020. Shape perception via a high-channel-count neuroprosthesis in monkey visual cortex. *Science*, 370(6521): 1191–1196.

Chen, Z.; Qing, J.; Xiang, T.; Yue, W. L.; and Zhou, J. H. 2023. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22710–22720.

Cogan, S. F. 2008. Neural stimulation and recording electrodes. *Annu. Rev. Biomed. Eng.*, 10(1): 275–309.

Dobelle, W.; and Mladejovsky, M. 1974. Phosphenes produced by electrical stimulation of human occipital cortex, and their application to the development of a prosthesis for the blind. *The Journal of physiology*, 243(2): 553–576.

Dobelle, W. H. 2000. Artificial vision for the blind by connecting a television camera to the visual cortex. *ASAIO journal*, 46(1): 3–9.

Dobelle, W. H.; Mladejovsky, M. G.; and Girvin, J. 1974. Artificial vision for the blind: electrical stimulation of visual cortex offers hope for a functional prosthesis. *Science*, 183(4123): 440–444.

Dorn, J. D.; Ahuja, A. K.; Caspi, A.; Da Cruz, L.; Dagnelie, G.; Sahel, J.-A.; Greenberg, R. J.; McMahon, M. J.; Group, A. I. S.; et al. 2013. The detection of motion by blind subjects with the epiretinal 60-electrode (Argus II) retinal prosthesis. *JAMA ophthalmology*, 131(2): 183–189.

Du, C.; Du, C.; Huang, L.; and He, H. 2018. Reconstructing perceived images from human brain activities with Bayesian deep multiview learning. *IEEE transactions on neural networks and learning systems*, 30(8): 2310–2323.

Farzmahdi, A.; Kohn, A.; and Coen-Cagli, R. 2025. Relating natural image statistics to patterns of response covariability in macaque primary visual cortex. *Nature Communications*, 16(1): 1–13.

Fernández, E.; Alfaro, A.; Soto-Sánchez, C.; Gonzalez-Lopez, P.; Lozano, A. M.; Peña, S.; Grima, M. D.; Rodil, A.; Gómez, B.; Chen, X.; et al. 2021. Visual percepts evoked with an intracortical 96-channel microelectrode array inserted in human occipital cortex. *The Journal of clinical investigation*, 131(23).

Fernandez, E.; and Robles, J. A. 2024. Advances and challenges in the development of visual prostheses. *PLoS biology*, 22(10): e3002896.

Fiorani, M.; Azzi, J. C.; Soares, J. G.; and Gattass, R. 2014. Automatic mapping of visual cortex receptive fields: a fast and precise algorithm. *Journal of neuroscience methods*, 221: 112–126.

Foroushani, A. N.; Pack, C. C.; and Sawan, M. 2018. Cortical visual prostheses: from microstimulation to functional percept. *Journal of neural engineering*, 15(2): 021005.

Gaziv, G.; Beliy, R.; Granot, N.; Hoogi, A.; Strappini, F.; Golan, T.; and Irani, M. 2022. Self-supervised natural image reconstruction and large-scale semantic classification from brain activity. *NeuroImage*, 254: 119121.

Grani, F.; Soto-Sánchez, C.; Fimia, A.; and Fernández, E. 2022. Toward a personalized closed-loop stimulation of the visual cortex: Advances and challenges. *Frontiers in Cellular Neuroscience*, 16: 1034270.

Granley, J.; Relic, L.; and Beyeler, M. 2022. Hybrid neural autoencoders for stimulus encoding in visual and other sensory neuroprostheses. *Advances in Neural Information Processing Systems*, 35: 22671–22685.

Hatanaka, G.; Inagaki, M.; Takeuchi, R. F.; Nishimoto, S.; Ikezoe, K.; and Fujita, I. 2022. Processing of visual statistics

of naturalistic videos in macaque visual areas V1 and V4. *Brain Structure and Function*, 227(4): 1385–1403.

Larkin, C. J.; Yerneni, K.; Karras, C. L.; Abecassis, Z. A.; Zhou, G.; Zelano, C.; Selner, A. N.; Templer, J. W.; and Tate, M. C. 2022. Impact of intraoperative direct cortical stimulation dynamics on perioperative seizures and afterdischarge frequency in patients undergoing awake craniotomy. *Journal of neurosurgery*, 137(6): 1853–1861.

Lewis, P. M.; Ackland, H. M.; Lowery, A. J.; and Rosenfeld, J. V. 2015. Restoration of vision in blind individuals using bionic devices: a review with a focus on cortical visual prostheses. *Brain research*, 1595: 51–73.

Li, B. M.; Cornacchia, I. M.; Rochefort, N. L.; and Onken, A. 2023. V1t: large-scale mouse v1 response prediction using a vision transformer. *arXiv preprint arXiv:2302.03023*.

Lozano, A.; Suárez, J. S.; Soto-Sánchez, C.; Garrigós, J.; Martínez-Alvarez, J. J.; Ferrández, J. M.; and Fernández, E. 2020. Neurolight: A deep learning neural interface for cortical visual prostheses. *International journal of neural systems*, 30(09): 2050045.

Luo, Y. H.-L.; and Da Cruz, L. 2016. The Argus® II retinal prosthesis system. *Progress in retinal and eye research*, 50: 89–107.

Maghami, M. H.; Sodagar, A. M.; Lashay, A.; Riazi-Esfahani, H.; and Riazi-Esfahani, M. 2014. Visual prostheses: The enabling technology to give sight to the blind. *Journal of ophthalmic & vision research*, 9(4): 494.

Oswalt, D.; Bosking, W.; Sun, P.; Sheth, S. A.; Niketeghad, S.; Salas, M. A.; Patel, U.; Greenberg, R.; Dorn, J.; Pouratian, N.; et al. 2021. Multi-electrode stimulation evokes consistent spatial patterns of phosphenes and improves phosphene mapping in blind subjects. *Brain stimulation*, 14(5): 1356–1372.

Ozcelik, F.; Choksi, B.; Mozafari, M.; Reddy, L.; and VanRullen, R. 2022. Reconstruction of perceived images from fmri patterns and semantic brain exploration using instance-conditioned gans. In *2022 international joint conference on neural networks (IJCNN)*, 1–8. IEEE.

Roelfsema, P. R.; Denys, D.; and Klink, P. C. 2018. Mind reading and writing: the future of neurotechnology. *Trends in cognitive sciences*, 22(7): 598–610.

Shipp, S. 2024. Computational components of visual predictive coding circuitry. *Frontiers in Neural Circuits*, 17: 1254009.

Sinz, F.; Ecker, A. S.; Fahey, P.; Walker, E.; Cobos, E.; Froudarakis, E.; Yatsenko, D.; Pitkow, Z.; Reimer, J.; and Tolias, A. 2018. Stimulus domain transfer in recurrent models for large scale cortical population prediction on video. *Advances in neural information processing systems*, 31.

Stevens, G. A.; White, R. A.; Flaxman, S. R.; Price, H.; Jonas, J. B.; Keeffe, J.; Leasher, J.; Naidoo, K.; Pesudovs, K.; Resnikoff, S.; et al. 2013. Global prevalence of vision impairment and blindness: magnitude and temporal trends, 1990–2010. *Ophthalmology*, 120(12): 2377–2384.

Stingl, K.; Bartz-Schmidt, K. U.; Besch, D.; Chee, C. K.; Cottriall, C. L.; Gekeler, F.; Groppe, M.; Jackson, T. L.; MacLaren, R. E.; Koitschev, A.; et al. 2015. Subretinal visual implant alpha IMS–clinical trial interim report. *Vision research*, 111: 149–160.

Tehovnik, E. J.; and Slocum, W. M. 2013. Electrical induction of vision. *Neuroscience & Biobehavioral Reviews*, 37(5): 803–818.

Tehovnik, E. J.; Slocum, W. M.; Carvey, C. E.; and Schiller, P. H. 2005. Phosphene induction and the generation of saccadic eye movements by striate cortex. *Journal of neurophysiology*, 93(1): 1–19.

Tehovnik, E. J.; Slocum, W. M.; and Schiller, P. H. 2003. Saccadic eye movements evoked by microstimulation of striate cortex. *European Journal of Neuroscience*, 17(4): 870–878.

Turishcheva, P.; Fahey, P.; Vystrčilová, M.; Hansel, L.; Froebe, R.; Ponder, K.; Qiu, Y.; Willeke, K.; Bashiri, M.; Baikulov, R.; et al. 2024a. Retrospective for the Dynamic Sensorium Competition for predicting large-scale mouse primary visual cortex activity from videos. *Advances in Neural Information Processing Systems*, 37: 118907–118929.

Turishcheva, P.; Fahey, P. G.; Vystrčilová, M.; Hansel, L.; Froebe, R.; Ponder, K.; Qiu, Y.; Willeke, K. F.; Bashiri, M.; Wang, E.; et al. 2024b. The dynamic sensorium competition for predicting large-scale mouse visual cortex activity from videos. *ArXiv*, arXiv–2305.

van der Grinten, M.; van Steveninck, J. d. R.; Lozano, A.; Pijnacker, L.; Rueckauer, B.; Roelfsema, P.; van Gerven, M.; van Wezel, R.; Güçlü, U.; and Güçlütürk, Y. 2024. Towards biologically plausible phosphene simulation for the differentiable optimization of visual cortical prostheses. *Elife*, 13: e85812.

van Steveninck, J. d. R.; Güçlü, U.; van Wezel, R.; and van Gerven, M. 2022. End-to-end optimization of prosthetic vision. *Journal of Vision*, 22(2): 20–20.

Vatsyayan, R.; Cleary, D.; Martin, J. R.; Halgren, E.; and Dayeh, S. A. 2021. Electrochemical safety limits for clinical stimulation investigated using depth and strip electrodes in the pig brain. *Journal of neural engineering*, 18(4): 046077.

Vatsyayan, R.; and Dayeh, S. A. 2022. A universal model of electrochemical safety limits in vivo for electrophysiological stimulation. *Frontiers in neuroscience*, 16: 972252.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Xu, A.; Hou, Y.; Niell, C.; and Beyeler, M. 2023. Multimodal deep learning model unveils behavioral dynamics of V1 activity in freely moving mice. *Advances in neural information processing systems*, 36: 15341–15357.

Zhang, Y.; Bu, T.; Zhang, J.; Tang, S.; Yu, Z.; Liu, J. K.; and Huang, T. 2022. Decoding pixel-level image features from two-photon calcium signals of macaque visual cortex. *Neural Computation*, 34(6): 1369–1397.

Zhang, Y.; Jia, S.; Zheng, Y.; Yu, Z.; Tian, Y.; Ma, S.; Huang, T.; and Liu, J. K. 2020. Reconstruction of natural visual scenes from neural spikes with deep neural networks. *Neural Networks*, 125: 19–30.

Appendix

A Detailed Architecture of MUA-Image Decoder (MID)

As shown in Figure 1 (in the main text), the overall calculation process of MID is as follows:

$$\hat{I}_i = \mathcal{D}_\psi(\mathcal{E}_\phi(\mathcal{P}_\theta(\text{MUA}_i)))$$

Here, $\hat{I}_i$ denotes the image decoded from MUA (1280 channels × 50 ms), $\mathcal{P}_\theta$ represents the projection network, and $\mathcal{E}_\phi$ and $\mathcal{D}_\psi$ denote the encoder and decoder of the Image-Image autoencoder, respectively.

Specifically, MUA is first processed through a spatiotemporal projection ($\mathcal{P}_\theta$). Each channel's waveform is processed using a one-dimensional convolution layer, followed by batch normalization (Ioffe and Szegedy 2015) and adaptive global temporal pooling to obtain the feature vector h. This feature is then passed through a fully connected layer and reshaped into an intermediate image h_{proj}. The detailed computation process is as follows:

$$h = \text{AvgPool}(\text{ReLU}(\text{BN}(\text{Conv1D}(\text{MUA}_i))))$$

$$h_{\text{proj}} = \mathcal{FC}(h), \in \mathbb{R}^{3\times 64\times 64}$$

The intermediate image representation is then fed into an autoencoder (Attention-Enhanced Encoder $\mathcal{E}_\phi$ combined with Biologically-Inspired Decoder $\mathcal{D}_\psi$) to reconstruct an image. The Attention-Enhanced Encoder performs four-stage feature extraction through an EncoderBlock composed of Conv2D → CBAM → MaxPool, capturing multi-scale and hierarchical refined features e_j from the intermediate image, where j represents the layer number of the feature. The detailed computation process is as follows:

$$e_1 = \text{CBAM}(\text{Conv2D}(h_{\text{proj}}))$$

$$e_j = \text{EncoderBlock}(e_{j-1}) \quad j = 2, 3, 4$$

To capture the selective tuning of visual cortical neurons to different visual attributes and the modulation of receptive fields, the Biologically-Inspired Decoder employs spatiotemporal attention to associate distinct parts and properties of the visual input with neural spiking activity across different channels and time points. To mimic the feedback and integration of information across visual cortical areas, the decoder further recombines multi-scale features through channel-wise concatenation with skip connections, enabling spatiotemporal fusion of multi-scale MUA dynamics. The detailed computation process is as follows:

$$d_4 = \text{ReLU}(\text{BN}(\text{Conv2D}(e_4)))$$

$$d_{j-1} = \text{CBAM}(\text{DecoderBlock}(\text{Cat}[\mathcal{U}(d_j), e_j]))$$

$$\hat{I} = \text{Tanh}(\text{Conv2D}(d_0))$$

Here, $\mathcal{U}$ consists of bilinear upsampling and Conv2D; DecoderBlock denotes (Conv2D→BN→ReLU)*2; e_j denotes the output of $\mathcal{E}_\phi$, which encodes multi-scale features of the intermediate image; Cat represents channel-wise concatenation with skip connections; and $j = 4, \ldots, 1$ indicates the feature level.

Mean squared error was calculated between output decoded images $\hat{I}$ and the original images I to train MID.

$$\mathcal{L}_{\text{MID}} = \mathbb{E}[\|I - \hat{I}\|_2^2]$$

B Implementation Details

Electrophysiological data is processed using the SpikeInterface and Mountainsort-5 toolbox (Chung et al. 2017; Buccino et al. 2020) to extract multi-unit activity (MUA) from the white-band signal. Specifically, the threshold for spike detection when applied to whitened data is 4.3 standard deviations. Subsequently, the MUA is downsampled from a 20kHz sampling frequency to 1kHz, and the 1280 channels with the highest firing rates are selected. Each 64×64 image corresponds to an MUA shape of 50×1280, as the MUA is taken over a 50ms period. We chose a 50ms MUA window as the decoding input for two main reasons. First, primary visual cortex exhibit initial response windows of 20–100ms, with neurons encoding maximal information within tens of milliseconds. Spike/PSTH counts show peak information entropy between 30–70ms. Second, guided by this biological basis, we narrowed the hyperparameter search range and ultimately identified 50ms as the optimal value. This data is used to train the MID, with a batch size of 32, a learning rate of 0.001, and the Adam optimizer to find the best model on the validation set over 1000 epochs, followed by testing on the test set.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Algorithm 1: Algorithmic workflow of the entire framework

Input: natural image I, multi-unit activity X, {impedance coefficient R_k, Spatial coordinates (x_k, y_k, z_k), center positions of the receptive fields $(\tilde{x}_k, \tilde{y}_k)$} of each electrode channel, labels Y indicating whether the image contains key information

Parameter: MUA-Image Decoder Θ_{mid}, Electrical Stimulation Generator ϕ_{esg}, Cortex-State Transition Process $\mathcal{B}$, Phosphene Simulator Ψ, Background Gated Network $\mathcal{F}$

Hyperparameters: Omitted for simplicity

Note: The following equation numbers all refer to those in the main text.

Output: Θ_{mid}, ϕ_{esg}, $\mathcal{F}$

randomly initialize Θ_{mid}, ϕ_{esg}, $\mathcal{F}$
while not done **do**
  Use the MUA-Image Decoder to obtain the decoded image $\hat{I}$, as shown in Equation (1)
  Compute decoding loss $\mathcal{L}_{\text{MID}} = \mathbb{E}[\|I - \hat{I}\|_2^2]$
  Perform gradient descent on $\mathcal{L}_{\text{MID}}$ to update Θ_{mid}
end while
For all I, generate binarized targets $\dot{I}$ through an adaptive double-threshold strategy
while not done **do**
  With Θ_{mid} fixed and $\mathcal{B}$ is detailed by Equations (6), (7), and (8), the decoded image $\hat{I}$ along Path 1 is obtained using Equation (5).
  With Ψ, the simulated phosphene image $\tilde{I}$ along Path 2 is obtained using Equation (10).
  Compute Path 1 loss $\mathcal{L}_{\text{Path1}} = \mathbb{E}[\|I - \hat{I}\|_2^2]$, Path 2 loss $\mathcal{L}_{\text{Path2}} = \mathbb{E}[\|\dot{I} - \tilde{I}\|_2^2]$, and the multi-channel activation constraint term as shown in Equation (12).
  As shown in Equation (13), combine losses from Path 1, Path 2, and channel activation constraint to obtain overall loss for Electrical Stimulation Generator ϕ_{esg}.
  Update ϕ_{esg} with gradient descent on overall loss
end while
while not done **do**
  Use Background Gated Network $\mathcal{F}$ to determine whether the image contains key information, as shown in Equation (14).
  Optimize $\mathcal{F}$ by minimizing the cross-entropy loss $\mathcal{L}_{\text{BGN}} = -\mathbb{E}\left[y \log p + (1-y)\log(1-p)\right]$.
end while
return Θ_{mid}, ϕ_{esg}, $\mathcal{F}$

The model parameters from the trained MID are then used to train the ESG, which involves numerous hyperparameters. Based on extensive impedance testing and statistical analysis across electrodes, we set the impedance coefficient R_k for each electrode channel to 10000. However, In practice, stimulation-induced impedance variations (influenced by factors such as electrode materials, tissue properties, and implantation duration) exhibit unpredictable complexity, currently excluded in modeling but prioritized for future work. During the cortex-state transition process, the number of pulses T_{pulse} is 1, the diffusion rate α is 2.0, and the initial standard deviation σ_0 is 5.0. For setting the value of α and σ_0, we first established a reasonable range based on (Wu et al. 2023), and then identified the optimal value through hyperparameter search. In the multi-channel activation constraint, the penalty coefficient γ is set to 5e-3, and the smoothing coefficient ϵ is set to 3e-6. For the loss summation, λ_1 is 0.3 and λ_2 is 0.7. During ESG training, the batch size is 32, the learning rate is 0.001, and the Adam optimizer is used to find the best model on the validation set over 1000 epochs, followed by testing on the test set.

During BGN training, the batch size is 32, the learning rate is 0.001, and the Adam optimizer is used to find the best model on the validation set over 50 epochs, followed by testing on the test set.

C Details of the dataset

1. **Monkey Neural-Image Viewing Dataset (NIVD)**: This dataset consists of neural signals and corresponding visual stimuli recorded while three monkeys (VM15, VM22, and VM23) viewed eight different shapes. The training set, validation set, and test set account for 80%, 10%, and 10%, respectively. 2. **Monkey Vehicle Obstacle Avoidance for Food — Neuro-Visual Dataset (MVOAF-NVD)**: This dataset includes neural signals and corresponding visual stimuli recorded while three monkeys (VM10, VM14, and VM23) drove a vehicle and avoided obstacles to obtain food in a sealed room. The dataset split is the same as that of NIVD. All experimental protocols involved in this study were approved and conducted in accordance with all relevant ethical regulations.

C.1 Details of Collecting the NIVD Dataset

The 1024-channel recording system (RHD2000, Intan Technologies) is connected to the equipment via a custom-designed adapter board. Each 128-channel headstage, containing two digital electrophysiology interface chips (Intan Technologies RHD2164, 64 channels), connects to the adapter board through a 64-pin Molex connector, with the ground wire soldered to the GND pin near index 127. We simultaneously recorded with 4 sets of devices to achieve 4096-channel recording. During the recording, the monkeys do not need to fixate on a central point (diameter 1.0°). In each trial, a shape stimulus is presented in the center of the gray screen for 1000ms. While the monkeys freely view the screen, a wearable eye-tracking system continuously captures the position of the eye's gaze center. The inter-trial interval is 3000ms, after which a new trial begins. The eight different shape stimuli (with randomly varying sizes) appear randomly in each session's trials. At the beginning of each trial, an event marker is sent to the recording system. The electrophysiological records and corresponding visual inputs are paired into NIVD samples through event markers and global time. Note: The visual input captures the 64 × 64-pixel region below the gaze center of the full screen at the corresponding moment (captured by the eye-tracking system). The 64 × 64 pixels are calculated based on the receptive field region covered by the implanted electrodes, screen resolution, and the distance between the monkey's eyes and

channels	electrical	visual
1	10.00%	97.50%
2	31.82%	97.56%
4	32.56%	97.62%
8	65.00%	95.24%
16	87.80%	95.00%

Table S1: Behavioral accuracy in the saccade-to-phosphene task (first column) and the visual saccade task (second column), with each row representing a different number of channels for electrical stimulation.

the screen, ensuring that this pixel region matches the receptive field region of the recording channels as closely as possible.

C.2 Details of Collecting the MVOAF-NVD Dataset

Supplementary video 1 (in Supplementary Material) provides an overhead perspective of a single trial in MVOAF-NVD. Using a recording system similar to NIVD, the monkey sits in a vehicle and uses a joystick to control its movement, avoiding obstacles in the room to reach the food dispenser and obtain food. Various shaped markers were attached to the walls and obstacles in the room. The visual input is obtained by a camera mounted on the monkey's head. Similarly, electrophysiological records and corresponding visual inputs are paired into MVOAF-NVD samples through event markers and global time.

D Method for Constructing Binarized Targets

In **Path 2: Phosphene Simulation Constraint**, for each real input I_i, we generate the corresponding binarized target $\dot{I}_i$ (with the foreground being the region that requires attention from electrical stimulation) through an adaptive double-threshold strategy: When color information is strong, it prioritizes segmentation of chromatic features to construct the binary mask; otherwise, the image is converted to grayscale, and the Otsu thresholding method (Otsu et al. 1975) is applied in combination with secondary percentile filtering.

E Electrode Implantation and Receptive Field Mapping

We used a fast RF mapping method similar with (Fiorani et al. 2014) to get the receptive fields of neurons. Figure S2 illustrates the electrode implantation and receptive field distributions for monkeys VM22 and VM23, respectively.

F Electrical-Stimulation-Induced Saccade

We followed previously reported methods (Chen et al. 2020) to conduct electrical-stimulation-induced saccade experiments on our monkeys. Specifically, we simultaneously stimulated varying numbers of channels on the same electrode to induce phosphenes. As shown in Table S1, by comparing task performance under visual stimulation versus electrical stimulation, we found that at least 16 channels on a single electrode must be stimulated simultaneously to achieve accuracy comparable to that under visual stimulation. This reflects the minimum number of channels required per electrode to elicit effective phosphene perception, which motivated our design of multi-channel activation constraint.

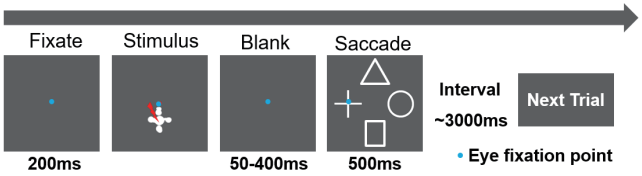

Figure S1: **llustration of the DMS Task.**

G Task Paradigm

We designed a delayed match-to-sample (DMS) task and a driving-based obstacle-avoidance foraging (DOAF) task to validate the efficacy of MindSight.

DMS. As illustrated in Figure S1, the monkey is first required to fixate on a central red dot (1.0°) for 200ms. Then, a randomly selected shape is fed into ESG to generate electrical stimulation parameters (executed by Intan RHS2116-based device). It is important to note that during this "stimulus" phase, the screen remains blank and the monkey's eyes are open. The "cross shape" shown under "stimulus" in Figure S1 is merely illustrative of the stimulation event and is not actually visible to the monkey. After stimulation, a 50–400ms delay (gray screen with a fixation point, and fixation must be maintained) preceded a selection phase: saccading to the matching shape among four options and holding fixation for 500ms. Inter-trial interval: 3000ms.

DOAF. Blindfolded monkeys navigated using a head-mounted camera feeding images to MindSight. When the BGN detected critical cues (e.g., obstacles, food dispensers), electrical stimulation triggered by the ESG enabled blindfolded monkeys to navigate obstacles and obtain food via cortical-stimulation-induced perception.

H Details of Supplementary Videos 2 and 3

The video is divided into four panels: the larger left section displays real-time screen footage, the upper-right panel shows camera-captured scenes, the right half of the lower-right panel shows tracking the monkey's gaze center via an eye tracker, and the left half of the lower-right panel simulates the visual perception evoked by the ESG's electrical stimulation commands (based on Equation (18) in the main text). Note that this simulated perception is invisible to the monkey and is included solely for visualization and analysis.

I Discussion

Performance validation in DOAF is confined solely to simulations of electrical stimulation due to the current electrode impedance issues; real-world electrical stimulation testing

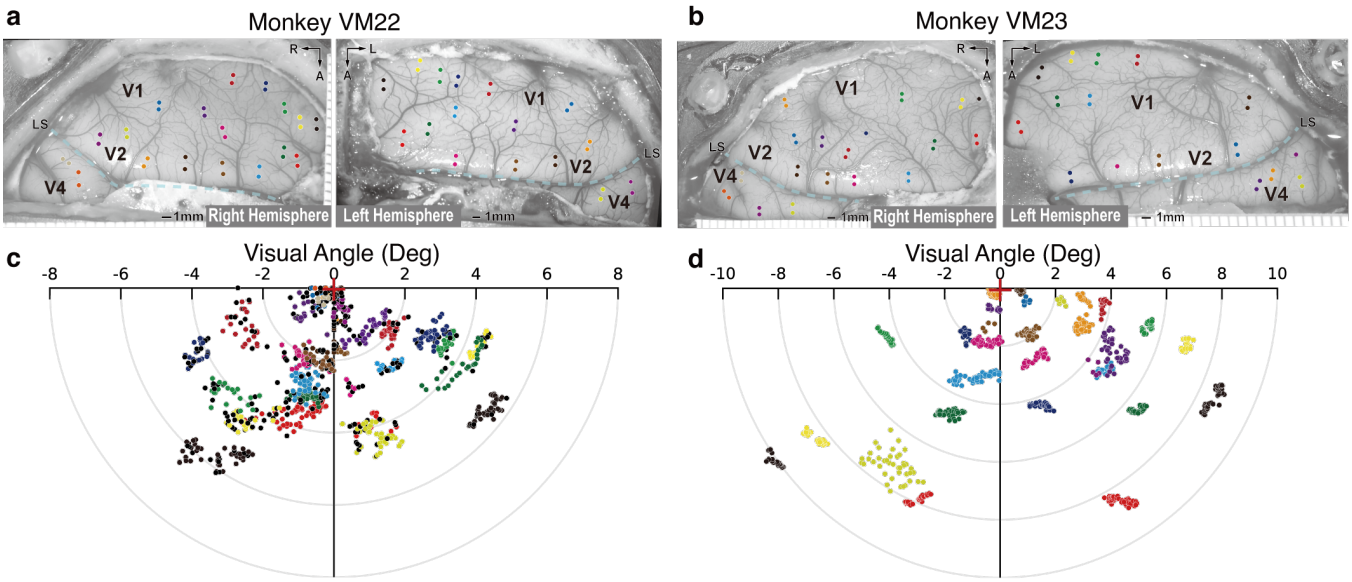

Figure S2: **Surgery Annotation and Receptive Fields. a** and **b** show surgery photograph of VM22 and VM23, respectively, overlaid with annotated brain regions and electrode implantation sites. Different colors represent different electrodes, and each electrode has two points because it consists of two shanks; each shank contains 64 channels. **c** and **d** show the receptive fields of VM22 and VM23, respectively (with the red cross indicating the center of each receptive fields), where each dot represents the receptive field location of a single channel from an implanted electrode (matched by color with **a** and **b**). Each electrode contains a total of 128 channels.

awaits re-implantation. Human validation remains challenging due to medical device constraints. Given the anatomical and perceptual similarities between macaques and humans, we used macaques as the subjects for this study. Our results demonstrate the potential of MindSight to propel the widespread application of visual prostheses in the future. Future work will focus on two main aspects: firstly, utilizing neural recordings obtained before and after cortical stimulation through simultaneous stimulation and recording techniques to study and optimize a more complex and realistic Cortex-State Transition Process; secondly, exploring how to dynamically adjust parameters during long-term stimulation to adapt to changes in biological states and electrode conditions.

J Data and Code Availability

You can find the data and code at https://github.com/zyj9902/MindSight.

References

Buccino, A. P.; Hurwitz, C. L.; Garcia, S.; Magland, J.; Siegle, J. H.; Hurwitz, R.; and Hennig, M. H. 2020. SpikeInterface, a unified framework for spike sorting. *Elife*, 9: e61834.

Chen, X.; Wang, F.; Fernandez, E.; and Roelfsema, P. R. 2020. Shape perception via a high-channel-count neuroprosthesis in monkey visual cortex. *Science*, 370(6521): 1191–1196.

Chung, J. E.; Magland, J. F.; Barnett, A. H.; Tolosa, V. M.; Tooker, A. C.; Lee, K. Y.; Shah, K. G.; Felix, S. H.; Frank, L. M.; and Greengard, L. F. 2017. A fully automated approach to spike sorting. *Neuron*, 95(6): 1381–1394.

Fiorani, M.; Azzi, J. C.; Soares, J. G.; and Gattass, R. 2014. Automatic mapping of visual cortex receptive fields: a fast and precise algorithm. *Journal of neuroscience methods*, 221: 112–126.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 448–456. pmlr.

Otsu, N.; et al. 1975. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296): 23–27.

Wu, G. K.; Ardeshirpour, Y.; Mastracchio, C.; Kent, J.; Caiola, M.; and Ye, M. 2023. Amplitude-and frequency-dependent activation of layer II/III neurons by intracortical microstimulation. *IScience*, 26(11).