PHOSINV: LEARNING A SHARED IMPLANT PLACEMENT AND STIMULUS-CONDITIONAL ELECTRODE SELECTION FOR VISUAL CORTICAL PROSTHESES

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual cortical prostheses aim to restore rudimentary vision by electrically stimulating the primary visual cortex to evoke percepts known as phosphenes. Their design requires determining where to implant a fixed electrode array and which contacts to activate for each target stimulus. Prior work reoptimises implant placement separately for each stimulus, an approach that cannot be implemented once a device has been implanted. We reformulate this task as a generalisation problem: learning a single implant placement shared across a distribution of target stimuli, together with a stimulus-conditional electrode-activation policy. PhosInv implements this formulation as a convolutional inverse model trained end-to-end through a differentiable reimplementation of an existing phosphene simulator. Once trained, it predicts a sparse electrode selection for each new stimulus in a single forward pass, without per-stimulus optimisation. On a benchmark of 2,820 targets from two domains, all scored under the differentiable renderer, domain-matched PhosInv attains lower loss than per-stimulus Bayesian placement search in all six subject hemispheres in both domains. A forward pass takes milliseconds on a GPU against minutes of CPU search per stimulus, a gap that combines amortisation with differences in hardware and implementation, and PhosInv commands six to eight times fewer contacts, reducing delivered charge, a quantity constrained by safety considerations. A model trained on an equal mixture of both domains retains most of each domain-specific model's advantage while avoiding the degradation either shows outside its own domain, indicating that training coverage, not the architecture itself, limits generalisation. Relative to a shared all-contacts-active reference, learned electrode selection contributes about three times the improvement achieved by per-stimulus placement search and far more than a placement searched once per hemisphere, identifying electrode selection as the more consequential lever for future high-density arrays. Every contrast is reported per anatomy, as an effect size and its consistency across six subject hemispheres.

1 INTRODUCTION

Visual cortical prostheses (VCPs) aim to restore rudimentary vision where retinal or optic-nerve damage rules out retinal implants or optogenetic therapy, by implanting an electrode array in the primary visual cortex (V1) and converting camera input into patterned stimulation that evokes localised percepts, or phosphenes. Clinical trials show such phosphenes support basic shape and motion perception (Beauchamp et al., 2020; Fernandez et al., 2021), but percept quality remains limited, partly because cortical magnification distorts the visual field (Horton & Hoyt, 1991) in ways that vary across individuals (Benson et al., 2018). An effective stimulation strategy must therefore decide both where an array is placed and which of its contacts are activated for a given target, and both depend on subject-specific anatomy.

Existing work addresses either half in isolation (Section 2): per-stimulus placement search re-optimises geometry for every target rather than for a placement a device could keep, and treats

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

selection as no part of the search (van Hoof et al., 2025); differentiable simulation has been applied only to a fixed, predefined geometry (van der Grinten et al., 2024); and learned encoders inverting a fixed forward simulator (de Ruyter van Steveninck et al., 2022; Granley et al., 2022; 2023) treat device geometry as a boundary condition. No framework optimises one shared placement across a distribution of stimuli together with a sparse, stimulus-conditional electrode selection, the pair of decisions a deployed device must make.

We reformulate VCP optimisation as generalisation over a target distribution rather than per-stimulus search (Section 3) and present PhosInv, a convolutional inverse model solving it end to end through a differentiable phosphene simulator (Section 4). Training determines one shared implant placement and the weights of a stimulus-conditional selection network together, by backpropagation through the simulator rather than by nested or alternating search; at inference the placement is fixed and a single forward pass predicts only the sparse electrode mask for the incoming stimulus.

Contributions.

- We reformulate VCP placement-and-selection optimisation as generalisation over a target distribution under a shared physical placement (equation 3), replacing independent per-stimulus search (equation 2), and derive PhosInv as its amortised solution.
- We construct a 2,820-target, two-domain benchmark and evaluate eleven methods through a single differentiable rendering pipeline.
- Under that renderer, domain-matched PhosInv attains lower loss than per-stimulus Bayesian search in every tested anatomy in both domains, with a single forward pass and a small fraction of the electrode budget, and a mixed-corpus model retains most of this advantage across both domains.
- Decomposing the two optimisation levers shows that which electrodes are activated matters substantially more than refining implant geometry, which matters as high-density arrays make selection combinatorially harder while placement stays low-dimensional.
- We report the moderate agreement between our two simulator implementations (rank correlation 0.55–0.61) and report every contrast at the anatomy level, as an effect size and a sign count over six units (Section 5.5).

2 Related work

Simulating cortical prosthetic vision. Phosphene simulators map a stimulation pattern and an implant geometry to the percept a recipient would report, using cortical magnification and receptive field structure to place and size each phosphene (van der Grinten et al., 2024). van Hoof et al. (2025) combined such a simulator with subject-specific 7T retinotopy, giving a forward model in which array placement can be evaluated against a desired percept. We adopt that forward model and make it differentiable.

Optimising implant placement. Placement has been treated as a black-box search problem, with Bayesian optimisation over insertion geometry run independently for each target stimulus (van Hoof et al., 2025). The resulting stimulus-conditional geometry is informative as a simulation study but cannot be realised by a device implanted once. We keep the same geometric parameterisation and search space and change what is optimised (Section 3).

Learning to invert a fixed forward model. Training an encoder jointly with a differentiable simulator was introduced for prosthetic vision by de Ruyter van Steveninck et al. (2022) and developed for sensory neuroprostheses by Granley et al. (2022; 2023), who learn patient-specific stimulation for an implant whose geometry is given, a fixed boundary condition. We move the geometry inside the learned problem, optimising placement and per-stimulus selection against one objective, with an explicit binary mask in place of a dense stimulation vector.

Amortised optimisation. Replacing per-instance optimisation with a trained predictor is a general pattern (Chen et al., 2022); here the shared variable is physically forced to be shared, so equation 3 is the correct target rather than an approximation to equation 2.

3 PROBLEM FORMULATION

Setup. Let a denote one subject hemisphere's anatomy: its retinotopic map and V1 grey matter segmentation. An implant is a placement vector $\vartheta \in \Theta \subset \mathbb{R}^4$: two insertion angles, an offset along the calcarine sulcus, and a shank length. Given ϑ and a the N contacts acquire fixed cortical positions, and each one inside V1 grey matter evokes a phosphene whose visual field position and size follow from the retinotopy of a. A binary mask $m \in \{0,1\}^N$ selects which contacts are stimulated, and a forward simulator $\mathcal{S}$ maps these to a predicted phosphene map,

$$P \;=\; \mathcal{S}(\vartheta, m; a) \;\in\; \mathbb{R}^{H\times W}, \tag{1}$$

and is differentiable with respect to ϑ and to a continuous relaxation of m. Reconstruction quality against a target map $T \in \mathbb{R}^{H\times W}$ is measured by the composite loss $\mathcal{L}(T, P)$ defined in Section 5.4.

Per-stimulus optimisation. Existing placement optimisation (van Hoof et al., 2025) treats each stimulus as an independent problem,

$$\vartheta^\star(T) \;=\; \arg\min_{\vartheta\in\Theta} \; \mathcal{L}(T,\; \mathcal{S}(\vartheta, m_{\mathrm{V1}}(\vartheta; a); a))\,, \tag{2}$$

where $m_{\mathrm{V1}}(\vartheta; a)$ activates every V1-intersecting contact and is recomputed for each candidate ϑ. The solution is a function of T, so a different stimulus implies a different implant geometry, and selection is not a decision variable at all. An implanted array is positioned once and cannot be repositioned between stimuli, so equation 2 is a per-stimulus lower envelope rather than an attainable operating point. Its deployable restriction, solving it for one reference target and applying the result unchanged to every subsequent stimulus, we call the fixed-reference construction and evaluate as an explicit baseline.

Generalisation objective. We replace it with the expected loss over a distribution $\mathcal{D}$ of target maps, under a placement shared across that distribution,

$$\min_{\vartheta\in\Theta,\; \pi_a\in\Pi} \; \mathbb{E}_{T\sim\mathcal{D}}\left[\mathcal{L}(T,\; \mathcal{S}(\vartheta, \pi_a(T); a))\right], \tag{3}$$

where $\pi_a : \mathbb{R}^{H\times W} \to \{0,1\}^N$ is a selection policy for anatomy a. Here ϑ is shared across all $T \sim \mathcal{D}$ while remaining specific to a, and only π_a adapts per stimulus; both are solved separately for each subject hemisphere. This mirrors the device: geometry fixed at implantation, stimulation updated every frame, and makes equation 3 a joint problem over a subject-level variable and a stimulus-level function rather than a restriction of equation 2.

Amortised solution. We parameterise π_a with a convolutional network f_ϕ and optimise ϑ and ϕ jointly by gradient descent through the simulator,

$$(\hat\vartheta, \hat\phi) \;=\; \arg\min_{\vartheta,\phi} \; \frac{1}{|\mathcal{D}_{\mathrm{train}}|} \sum_{T\in\mathcal{D}_{\mathrm{train}}} \mathcal{L}(T,\; \mathcal{S}(\vartheta, f_\phi(T); a))\,. \tag{4}$$

Differentiability of $\mathcal{S}$ lets the gradient reach ϑ through the rendering geometry and ϕ through the predicted mask in one backward pass. Equation 4 is paid once per subject hemisphere and inference is a forward pass, whereas equation 2 is paid on every query and never amortises (Section 6.4).

Two optimisation levers. The two equations differ along two axes: placement per-stimulus or shared, and selection fixed or learned. This gives the design Section 6.3 decomposes. Only the cell combining per-stimulus placement with learned selection is unattainable: it needs a combinatorial selection search nested inside every placement optimisation.

Why not search placement first. A decoupled alternative also targets equation 3: search ϑ once per hemisphere, then learn π_a on top. It judges ϑ under m_{V1} rather than under the policy that will drive it, and needs a separate search stage; solving equation 4 jointly selects ϑ for the policy it serves, in one run. Section 6.3 compares the learned placement with the first stage of that alternative (the fixed-reference placement) under identical all-on selection. The full two-stage pipeline is not evaluated, so we make no claim that joint learning is superior to the decoupled pipeline.

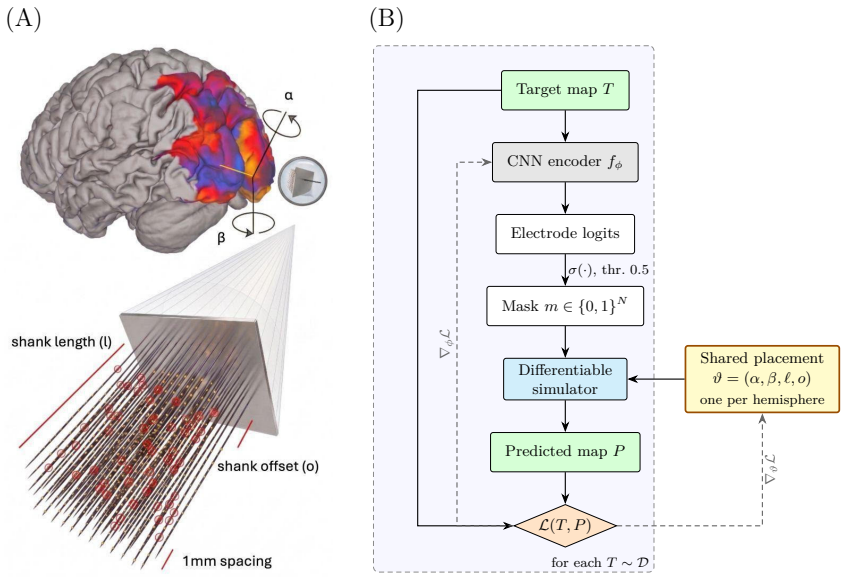

Figure 1: Joint optimisation of implant placement and electrode selection. **(A)** Subject-specific V1 retinotopy on a reconstructed cortical surface, with the array trajectory parameterised as in Section 3. **(B)** The inverse model. Everything inside the dashed plate is recomputed for every target; ϑ sits outside it as one parameter shared across the distribution, entering the simulator directly rather than through the encoder. The loss is backpropagated to the encoder weights and, across the plate boundary, to ϑ (equation 4).

Dependence on the target distribution. A solution of equation 3 is defined relative to $\mathcal{D}$, so we instantiate a letter-like $\mathcal{D}_{\mathrm{L}}$, a geometric $\mathcal{D}_{\mathrm{G}}$ and their equal mixture $\mathcal{D}_{\mathrm{M}}$, making training coverage an explicit variable (Section 6.2).

4 Method

4.1 Differentiable phosphene simulator

The forward simulator $\mathcal{S}$ builds on the phosphene model of van Hoof et al. (2025): a $10 \times 10 \times 10$ grid of $N = 1000$ contacts is positioned by $\vartheta = (\alpha, \beta, \mathrm{offset}, \mathrm{shank})$, each contact inherits pRF properties from nearby V1 voxels, and each is rendered as a Gaussian on the 128×128 visual field map with spread set by the cortical magnification factor at that eccentricity. That implementation is NumPy based and relies on discrete voxel lookup and non-differentiable ray-surface intersection, which precludes gradient-based optimisation of ϑ.

We reimplemented $\mathcal{S}$ in PyTorch with three changes (interpolated surface distances, a Gaussian-weighted rather than nearest-voxel assignment of receptive field properties, and a continuous relaxation of m during training) that together make $\vartheta \mapsto \mathcal{S}(\vartheta, m; a)$ differentiable in both arguments, which is what allows equation 4 to be solved by backpropagation rather than black-box search. Appendix B gives the equations of each change and of the rendering operator. The phosphene size scale factor is held at the default $s = 0.5$ of van der Grinten et al. (2024) throughout and never fit to reconstruction data; a prospective, unvalidated calibration of this factor for future benchmark use is sketched in Appendix N.

4.2 Inverse model architecture and training objective

The objective equation 4 has two variables, ϑ and ϕ, but only the second is a network: f_ϕ instantiates the policy π_a of equation 3 as a convolutional encoder with a single electrode head. The encoder maps a 128×128 target through three downsampling stages and four residual blocks to a 256-dimensional latent vector, following de Ruyter van Steveninck et al. (2022); the electrode head is

three fully connected layers giving $N = 1000$ logits, passed through a sigmoid for the continuous relaxation used in equation 4 and thresholded at 0.5 at inference (Appendix B).

The placement ϑ is *not* produced by the encoder: it is one learnable parameter in $\mathbb{R}^4$, shared across every training example in a subject hemisphere, initialised at the midpoint of its range and mapped into it by a sigmoid. Because it does not depend on the input, its gradient accumulates across the batch and it is optimised for the training distribution as a whole, while the electrode head varies per stimulus by construction; this alone enforces the sharing of equation 3, without any regularisation term.

Training minimises equation 4 with the form of equation 5 (Section 5.4) but with a soft Dice $\mathrm{DC}_{\mathrm{soft}}$ computed on the map rendered from the continuous sigmoid mask, so the loss is differentiable with respect to the electrode logits; Y likewise comes from the simulator's continuous validity scores. Evaluation applies equation 5 with DC on the binarised prediction. The two are not interchangeable and can order the same two conditions differently even when neither is in error, which Section 6.5 uses directly. All models are optimised with AdamW, a higher learning rate on ϑ than on the encoder, and early stopping on validation loss (Appendix B).

4.3 Training regimes

Three instances of equation 4 are trained per subject hemisphere, differing only in $\mathcal{D}_{\mathrm{train}}$. *PhosInv-letter* and *PhosInv-geometric* train on $\mathcal{D}_{\mathrm{L}}$ and $\mathcal{D}_{\mathrm{G}}$, each a domain-specific corpus of 112,800 maps split 80/10/10. *PhosInv-mixed* trains on the equal mixture, stride-sampling 56,400 maps from each corpus for the same total of 112,800, so compute is matched and, with architecture, optimiser and schedule shared, training coverage is the sole independent variable in Section 6.2.

5 Experimental setup

5.1 Subject data

Experiments use the HCP 7T Retinotopy dataset (Benson et al., 2018). Cortical surfaces were reconstructed with FreeSurfer, empirical retinotopic maps warped into each subject's native space, and Bayesian retinotopic inference (Benson & Winawer, 2018) applied to obtain polar angle, eccentricity, receptive field size and visual area labels in volume space; the calcarine sulcus centroid gives the reference insertion point (Appendix A).

Benchmarks run on three subjects and both hemispheres, giving six subject-hemisphere units, each a separate optimisation problem with its own placement vector and, for the learned methods, its own model; we make no claim about anatomical variability across individuals. Scale is set by the per-stimulus baselines, which need one full optimisation run for each of $2{,}820 \times 6 = 16{,}920$ target-by-unit pairs, and we prioritised targets over subjects because the central claims concern generalisation across stimuli (Section 7).

5.2 Benchmark targets

Targets are 128×128 normalised intensity maps under a circular right hemifield mask. The letter-like domain $\mathcal{D}_{\mathrm{L}}$ derives from EMNIST Balanced (Cohen et al., 2017) (47 classes); the geometric domain $\mathcal{D}_{\mathrm{G}}$ consists of arc sectors and Gaussian activation patterns sampled over position, scale and count, covering contiguous fills and radially symmetric structures letters lack. Evaluation uses 2,820 targets, 1,410 per domain, including four canonical arcs from van Hoof et al. (2025) held fixed for comparability. Training and validation corpora are built separately per domain (90,240 and 11,280 maps), disjoint from evaluation (Appendix C). This replaces an earlier 30-target benchmark whose difficulty tiers were defined by the learned model's own loss.

5.3 Compared methods

Eleven methods span the three evaluation conditions of Section 3. **Per-stimulus:** *BO* (Bayesian optimisation) reimplements the placement search of van Hoof et al. (2025), solving equation 2 independently for every target by Gaussian process optimisation with all V1-intersecting contacts active;

Overlap starts from BO's placement and additionally selects contacts whose rendered phosphene centroid falls on target intensity above a fixed threshold, giving per-stimulus selection without learning. Both report the full search cost. **Fixed-reference:** *BO (fixed-ref.)* and *Overlap (fixed-ref.)* solve the same problems for one reference target and apply the result unchanged to every stimulus, so the predicted map is identical across targets. **Shared-placement adaptive:** the three training regimes of Section 4.3; *All-on*, which keeps PhosInv's placement but activates every V1-intersecting contact and so ablates the selection head with architecture, pipeline and placement held fixed; and *Random*, *Intensity* and *Center*, which activate k contacts matched to PhosInv's mean active count, controlling for electrode count while varying only the rule. Reference-target selection, search settings, thresholds and hyperparameters are in Appendices B and G.

5.4 Metrics and rendering pipeline

Reconstruction quality follows van Hoof et al. (2025). Dice DC measures spatial overlap between target and prediction, each normalised to unit peak and binarised at the fixed intensity 0.05 over the full 128×128 map. Yield Y is the proportion of the $N = 1000$ contacts intersecting V1 grey matter; since it depends only on ϑ, it is identical for every method sharing a placement (Appendix F). Hellinger distance HD measures dissimilarity between the normalised intensity distributions. The composite loss is

$$\mathcal{L} \;=\; 2 - \mathrm{DC} - 0.1\,Y + \mathrm{HD}, \tag{5}$$

with $\mathrm{DC}, Y, \mathrm{HD} \in [0, 1]$ and so $\mathcal{L} \in [0.9, 3]$, weighted as in van Hoof et al. (2025) for comparability; wall-clock cost and the number of commanded contacts are recorded alongside, the latter because delivered charge scales with the number of stimulated sites. The training objective's soft Dice and the binarised equation 5 can order conditions differently (Section 6.5); all reported quantities use the binarised form.

Single rendering pipeline. Every predicted map entering equation 5 is rendered through the same differentiable simulator, a standardisation that keeps rendering implementation from being confounded with electrode configuration. It is not neutral: BO searched against the reference NumPy implementation of van Hoof et al. (2025), which agrees only moderately with ours (Section 6.5), while PhosInv is trained through the differentiable one. Every comparison is therefore a claim under the differentiable renderer, not under the reference one.

5.5 Statistical analysis

Observations are non-independent along two axes: targets within a subject hemisphere share one placement and one model, and the same 2,820 targets recur across methods. The fully crossed mixed-effects model this design calls for did not converge reliably at this scale and nothing is reported from it. We therefore reduce every contrast to the six subject-hemisphere means and report, as primary evidence, the mean difference, the number of units sharing its sign, and the range of the per-unit differences. Six units from three subjects are not fully independent and support little formal inference: cluster-robust standard errors are liberal at this count, and the cluster Wilcoxon test cannot go below $p = 0.031$. Both p-values, raw and Holm-corrected, are in Appendix J as secondary evidence, and no claim rests on either. The two domains are analysed separately throughout, and all runs use a fixed random seed.

6 Results

6.1 Benchmark comparison

Table 1 reports mean composite loss by evaluation domain under the single rendering pipeline. The headline contrast is per-stimulus search against the domain-matched inverse model. On geometric targets PhosInv-geometric attains 1.741 against BO's 2.025, a mean difference of -0.284 that has the same sign in all six subject hemispheres (per-unit range -0.491 to -0.173); on letter targets PhosInv-letter attains 1.764 against 2.190, a difference of -0.426, again 6/6 (range -0.632 to -0.250). Both are claims under the differentiable renderer, which BO's search did not optimise against (Section 5.4).

Table 1: Mean composite loss (equation 5) by evaluation domain under the single rendering pipeline of Section 5.4; $n = 8{,}460$ per domain per method. Active ratio is the mean fraction of the $N = 1000$ contacts commanded on; BO and All-on command all of them, of which only the V1-intersecting ones evoke a phosphene (Appendix K). The other five methods are in Appendix D and Table 8.

Method	Loss (geo.)	Loss (letter)	Dice (geo./letter)	Active ratio (geo./letter)
BO (per-target)	2.025	2.190	0.471 / 0.425	1.00 / 1.00
Overlap (per-target)	2.339	2.297	0.264 / 0.332	0.23 / 0.14
PhosInv-letter	1.917	1.764	0.549 / 0.655	0.18 / 0.12
PhosInv-geometric	1.741	2.134	0.642 / 0.456	0.17 / 0.07
PhosInv-mixed	1.801	1.810	0.622 / 0.631	0.21 / 0.13
All-on	2.170	2.389	0.395 / 0.303	1.00 / 1.00

This does not imply per-target dominance. Pooled over the six hemispheres, PhosInv attains lower loss on 95.2% of letter and 93.0% of geometric targets, with per-hemisphere rates spanning 88.0–99.9% and 85.7–99.9%, so the advantage is large in aggregate but not uniform across anatomies. Figure 3 in Appendix D shows eight reconstructions, two of them cases where BO is stronger.

Overlap, the non-learned per-stimulus selector, does not close the gap: it attains *higher* loss than BO in both domains (2.339 and 2.297 against 2.025 and 2.190), so restricting BO's all-on activation to a spatially-matched subset degrades reconstruction once every method is scored under one renderer. PhosInv exceeds Overlap by a wider margin than it exceeds BO in both domains (-0.598 geometric, -0.532 letter, 6/6 in each), so learned selection leads even against a spatially-informed non-learned selector, not only against the naïve count-matched heuristics of Appendix L.

The fixed-reference constructions are uniformly worse than their per-stimulus counterparts (2.133/2.418 against 2.025/2.190 for BO; 2.429/2.513 against 2.339/2.297 for Overlap, geometric/letter), confirming a single frozen reconstruction is a materially weaker baseline than per-stimulus search. Three of the four comparisons hold in six or five of six hemispheres; the exception, BO versus BO (fixed-ref.) on geometric targets (-0.108), holds in four of six and is not read as established (per-unit values and secondary p-values in Table 3, Appendix D).

6.2 Cross-domain transfer

Evaluating the three training regimes on both domains isolates training-distribution coverage as the sole independent variable (Table 6, Appendix H). Domain matching holds asymmetrically. On letter targets it is unambiguous: PhosInv-letter beats PhosInv-geometric by -0.370 in all six hemispheres (1.764 versus 2.134), while PhosInv-mixed and PhosInv-letter differ by only 0.045 with the sign split 4/6: the deficit is specific to training *exclusively* outside the letter domain, not to a mixture that includes it. On geometric targets the mean favours PhosInv-geometric over PhosInv-letter (1.741 versus 1.917) but the direction holds in only three of six hemispheres, so we do not read it as an established effect.

PhosInv-mixed sits between the specialists (1.801 geometric, 1.810 letter), and its difference from either matched specialist is small and inconsistent in sign (3/6 and 4/6). For a deployable single model the informative comparison is against the *mismatched* specialist, since a device cannot know in advance which stimuli it will face; there PhosInv-mixed wins in every hemisphere in both domains, beating PhosInv-letter on geometric targets (1.801 versus 1.917, -0.116) and PhosInv-geometric on letter targets (1.810 versus 2.134, -0.325), both 6/6. One mixed-corpus model thus recovers most of each specialist's advantage while avoiding the degradation either suffers off its own domain.

6.3 Decomposition of the two optimisation levers

The comparisons so far differ in two respects at once. This section separates them using the placement-by-selection design of Section 3, reading each lever's marginal effect against the shared All-on reference, which keeps PhosInv's placement but activates every V1-intersecting contact, holding architecture, pipeline and placement fixed while removing only the selection head.

Against that reference (2.389 letter, 2.170 geometric), learned selection is worth 0.624 and 0.428 (PhosInv 1.764/1.741) while per-stimulus placement search is worth 0.199 and 0.144 (BO 2.190/2.025), ratios of 3.1 and 3.0, with the mixed-corpus model giving 0.579 and 0.368 in the same selection cell (Table 7 and Figure 6, Appendix I). The ordering holds for both the letter-matched and mixed-corpus models, so it is not an artefact of one training distribution: once a placement giving adequate V1 coverage has been found, which electrodes are activated matters considerably more than further refinement of the geometry.

The placement lever, against the right comparator. Holding selection at all-on, BO's per-stimulus placement beats the learned one by 0.199 on letter (6/6) and 0.144 on geometric targets (5/6), but that sets a shared placement against one refitted to each stimulus, which no implant can realise, so it measures the price of sharing. The deployable comparator is a placement searched once, BO (fixed-ref.), under the same all-on selection. Mean differences are small against the selection effects (learned better by 0.029 on letter, searched better by 0.037 on geometric), but not even across units: letter splits 3/6 (range -0.336 to 0.125), while on geometric targets the searched placement is better in five of six hemispheres (range -0.504 to 0.312). The learned placement is thus not uniformly as good as a search-once one, although ϑ was optimised for the learned mask rather than for all-on activation (Appendix I). What the decomposition establishes is that the gain over search comes from the selection head, a property of the selection rule rather than of sparsity (Appendix L); whether selection learned on a searched placement would do better is not evaluated (Section 3), and the two effects are marginal rather than additive.

What the selection lever consists of. With the learned placement fixed, PhosInv attains the lowest loss of five selection rules in both domains, and all three count-matched heuristics are *worse* than All-on despite activating six to eight times fewer contacts, on Dice alone as well: what matters is which contacts are activated, not how many (Appendix L).

6.4 Inference cost and amortisation

Equation 2 must complete an optimisation run before any stimulus can be presented; the amortised model must not. A forward pass of f_ϕ plus rendering takes 7.3–11.1 ms (90–137 Hz) on one RTX 5070 GPU, against 555.4 s per BO query run on a CPU host against the reference NumPy simulator. This ratio of about four and a half orders of magnitude is not a matched speedup: it combines amortisation with the GPU and the PyTorch reimplementation, which we did not separate. Only the structural difference is hardware-independent: one optimisation run per stimulus against one forward pass. We claim compatibility with real-time operation, not a perceptual update rate. Training costs 72,768 s (20.2 h, GPU) once per subject hemisphere, so under these conditions the totals cross at 131 stimuli, a figure inheriting the same hardware difference. The domain-matched model commands 170.5 contacts per stimulus on geometric and 124.6 on letter targets of the $N = 1000$ available, against all 1000 for BO and All-on (Table 1), lowering delivered charge at equal or better reconstruction quality (Appendix K).

6.5 Robustness

Four checks bound the contrasts; Appendix M gives each in full. Resampling one third of the letter evaluation set 500 times preserves the sign of both headline contrasts in 500 of 500 draws. All six fixed-reference variants are worse than the domain-matched model in their own domain. The two simulator implementations agree only moderately on individual configurations (Dice rank correlation 0.546, loss 0.611), which is why every method is scored under one of them; since each method family was fitted to a different renderer, part of the gap may reflect that fit, and a full re-scoring under the reference implementation is not reported. An earlier claim that yield accounts for roughly 30% of the PhosInv-minus-BO gap does not survive recomputation (2.1–3.2%) and is retracted.

7 Discussion and limitations

Why an amortised model beats instance-specific search. Bounded per-stimulus search cannot exploit regularities recurring across a distribution, whereas the inverse model is trained on all of it at once, at the cost of the placement freedom search retains.

Selection as the primary design target. For future high-density arrays, where selection combinatorics grow with channel count while placement stays low-dimensional, the activation policy is both the more consequential design target and the one more amenable to being learned rather than searched.

Cross-domain transfer as a training-coverage property. Nothing in Equation 3 bounds the breadth of $\mathcal{D}$, and PhosInv-mixed shows that widening it recovers most of each specialist's advantage without trading away the within-distribution result, so what limits generalisation here is training coverage, not the reformulation.

Limitations. Under all-on selection the learned placement is close to a search-once placement on average but not uniformly as good, being worse in five of six hemispheres on geometric targets (Section 6.3), and a learned selection head on top of a searched placement is not evaluated. We therefore do not establish that learning ϑ jointly is necessary or better than searching it first; that comparison is the first experiment we would add. All comparisons are made under the differentiable renderer, which agrees only moderately with the reference implementation (Section 6.5). Six anatomical units from three subjects support effect sizes and sign counts rather than formal population inference, and do not characterise anatomical variability across a patient population. All metrics are simulation-based, their relation to perceived quality in a recipient is unestablished, and the pipeline requires 7T fMRI retinotopy, unavailable in blind recipients; a structural-MRI retinotopy predictor (Ribeiro et al., 2021) could remove that dependency, though such predictors are trained on sighted cortex. Stronger learned encoders (Granley et al., 2022) and natural-scene or clinical-acuity targets are not evaluated; Appendix E discusses these and further limitations.

8 Conclusion

We reformulated visual cortical prosthesis optimisation as learning one implant placement shared across a target distribution together with a stimulus-conditional activation policy, rather than searching for both independently per stimulus. Under the differentiable renderer this strictly more constrained problem costs no reconstruction quality within the training distribution: PhosInv attains lower loss than per-stimulus Bayesian search in every subject-hemisphere anatomy tested, with a forward pass in place of a search and on a small fraction of the contacts, and what bounds the approach's breadth is training coverage rather than architecture. The gain comes from learned selection rather than from the learned placement, and learning selection on top of a placement searched once per hemisphere is the experiment we would run next.

AI use statement

In this work we used generative AI (large language model) tools for tasks carrying a required disclosure under the ICLR 2027 AI policy, specifically: implementing experimental code (the differentiable simulator reimplementation, the inverse model training and inference pipelines, the benchmark harness, and the statistical analysis scripts), providing feedback on experimental methodology (including flagging the initial mixed-effects model's convergence failure, described in Section 5.5, and the renderer-mismatch issue described in Section 6.5), and interpreting experimental results (translating recorded metric outputs into the comparisons and effect sizes reported in Section 6). We additionally used generative AI tools for tasks carrying a recommended disclosure: drafting and editing sections of the manuscript text, creating and editing analysis and figure-generation code, formatting references, and proposing the paper's section structure. Generative AI tools were not used to generate the original research idea or the reformulation of Section 3, which predate their involvement in this project, and were not used to select which findings to report; negative or inconvenient results (Section 6.2's asymmetric domain matching, the retracted yield-contribution claim in Section 6.5, and the unexplained left-hemisphere disadvantage in Appendix E) are reported because they were found in the recorded experimental outputs, not because of any editorial choice by the tools used to help draft the text describing them. All AI-assisted code was read, executed, and checked against its intended behaviour, and against independent recomputation of key contrasts, before any result derived from it was reported; all AI-assisted text was reviewed and revised by the authors. Numerical values quoted in the paper were taken from the recorded outputs of these runs and cross-checked against the source tables rather than reproduced from any model-generated

summary. We take responsibility for the final content of this work, including its text, claims, and artefacts.

Ethics statement

This work uses a publicly available, de-identified human neuroimaging dataset under its published terms of use. No new data were collected from human participants and no intervention of any kind was performed. All results are obtained from simulation. The reconstruction metrics reported here measure agreement between a simulated percept and a target image, and they are not measures of perceptual quality, functional vision, or clinical benefit. Nothing in this work should be read as evidence that the described configurations are safe or effective in a recipient, and the step from simulated reconstruction quality to perceptual outcome remains unvalidated. We note in particular that the reduction in the number of activated contacts reported here lowers delivered charge in the model only, and that safety of any stimulation protocol would require separate empirical assessment. The population from which the retinotopic data are drawn is sighted, whereas the intended recipients of a cortical prosthesis are not, and this mismatch is discussed in Section 7.

Reproducibility statement

Section 3 states the objectives being optimised and fixes the notation used throughout. Section 4 describes the simulator modifications, the model, and the training objective, with full layer specifications, optimiser settings, and search hyperparameters in Appendix B. Section 5 specifies the subject data, the construction of the evaluation and training corpora, every compared method, the metric definitions, and the statistical procedure, with the anatomical preprocessing pipeline in Appendix A and the target construction in Appendix C. All runs use a fixed random seed, stated in Appendix B. Section 6.5 reports the checks we ran on the stability of the reported contrasts, including agreement between the two simulator implementations, and Appendix D contains the supporting tables.

An anonymised code archive is submitted as supplementary material. It contains the differentiable simulator, the inverse model, all training, inference and benchmark scripts, the target-generation code, the anatomical preprocessing pipeline of Appendix A, and the per-target result tables from which every number in this paper is computed, together with a mapping from each table and figure to the script that produces it. Every table and figure in the main text can be regenerated from the included result tables without a GPU. The HCP retinotopy inputs and the EMNIST-derived training corpus are not redistributed (the former under the dataset's terms of use, the latter for size), and the archive gives the pipelines that reconstruct both. The finite-element and compartmental-neuron outputs behind Appendix N are likewise not included; that calibration is not applied to any result reported here.

References

Michael S. Beauchamp et al. Dynamic stimulation of visual cortex produces form vision in sighted and blind humans. *Cell*, 181:774–783.e5, 2020. doi: 10.1016/j.cell.2020.04.033.

Noah C. Benson and Jonathan Winawer. Bayesian analysis of retinotopic maps. *eLife*, 7:e40224, 2018. doi: 10.7554/eLife.40224.

Noah C. Benson, Keith W. Jamison, Michael J. Arcaro, Alexander T. Vu, Matthew F. Glasser, Timothy S. Coalson, David C. Van Essen, Essa Yacoub, Kamil Ugurbil, Jonathan Winawer, and Kendrick Kay. The Human Connectome Project 7 Tesla retinotopy dataset: Description and population receptive field analysis. *Journal of Vision*, 18(13):23, 2018. doi: 10.1167/18.13.23.

Nicholas T. Carnevale and Michael L. Hines. *The NEURON Book*. Cambridge University Press, Cambridge, 2006. doi: 10.1017/CBO9780511541612.

Tianlong Chen, Xiaohan Chen, Wuyang Chen, Howard Heaton, Jialin Liu, Zhangyang Wang, and Wotao Yin. Learning to optimize: A primer and a benchmark. *Journal of Machine Learning Research*, 23(189):1–59, 2022.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: Extending MNIST to handwritten letters. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pp. 2921–2926, 2017. doi: 10.1109/IJCNN.2017.7966217.

Jaap de Ruyter van Steveninck, Umut Güçlü, Richard van Wezel, and Marcel van Gerven. End-to-end optimization of prosthetic vision. *Journal of Vision*, 22(2):20, 2022. doi: 10.1167/jov.22.2.20.

Eduardo Fernandez et al. Visual percepts evoked with an intracortical 96-channel microelectrode array inserted in human occipital cortex. *Journal of Clinical Investigation*, 131:e151331, 2021. doi: 10.1172/JCI151331.

Jacob Granley, Lucas Relic, and Michael Beyeler. Hybrid neural autoencoders for stimulus encoding in visual and other sensory neuroprostheses. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22671–22685, 2022. doi: 10.52202/068431-1647.

Jacob Granley, Tristan Fauvel, Matthew Chalk, and Michael Beyeler. Human-in-the-loop optimization for deep stimulus encoding in visual prostheses. In *Advances in Neural Information Processing Systems*, volume 36, pp. 79376–79398, 2023. doi: 10.52202/075280-3474.

Jonathan C. Horton and William F. Hoyt. The representation of the visual field in human striate cortex. *Archives of Ophthalmology*, 109:816–824, 1991. doi: 10.1001/archopht.1991.01080060080030.

Zachary F. Mainen and Terrence J. Sejnowski. Influence of dendritic structure on firing pattern in model neocortical neurons. *Nature*, 382:363–366, 1996. doi: 10.1038/382363a0.

Fernanda L. Ribeiro, Steffen Bollmann, and Alexander M. Puckett. Predicting the retinotopic organization of human visual cortex from anatomy using geometric deep learning. *NeuroImage*, 244:118624, 2021. doi: 10.1016/j.neuroimage.2021.118624.

Maureen van der Grinten, Jaap de Ruyter van Steveninck, Antonio Lozano, Laura Pijnacker, Bodo Rueckauer, Pieter Roelfsema, Marcel van Gerven, Richard van Wezel, Umut Güçlü, and Yağmur Güçlütürk. Towards biologically plausible phosphene simulation for the differentiable optimization of visual cortical prostheses. *eLife*, 13:e85812, 2024. doi: 10.7554/eLife.85812.

Rick van Hoof, Antonio Lozano, Feng Wang, P. Christiaan Klink, Pieter R. Roelfsema, and Rainer Goebel. Optimal placement of high-channel visual prostheses in human retinotopic visual cortex. *Journal of Neural Engineering*, 22:026016, 2025. doi: 10.1088/1741-2552/adaeef.

A Anatomical preprocessing

Data come from the HCP 7T Retinotopy dataset (Benson et al., 2018), acquired on a Siemens 7T Magnetom scanner with a 32-channel receiver coil at 1.6 mm isotropic resolution and 1 s TR. T1- and T2-weighted structural scans, acquired at 0.7 mm and resampled to the 1.6 mm isotropic grid of the 7T pipeline, were processed with FreeSurfer 7.2.0 (`recon-all -T2pial -all`) to reconstruct the cortical surface. Empirical pRF maps (polar angle, eccentricity, receptive field size, R^2) distributed via the OSF repository of Benson & Winawer (2018) were warped from `fsaverage` to each subject's native surface space with `mri_surf2surf`, and Bayesian retinotopic inference was performed with Neuropythy's `register_retinotopy`, producing four volume-space maps per subject: `inferred_angle.mgz`, `inferred_eccen.mgz`, `inferred_sigma.mgz` and `inferred_varea.mgz`.

Together with `T1.mgz` and `aparc+aseg.mgz` these are the six input files the simulator requires, and no other subject-specific input enters the pipeline anywhere. The `aparc+aseg.mgz` parcellation supplies the calcarine sulcus centroid used as the reference insertion point and identifies grey matter and V1 voxels; `T1.mgz` defines the coordinate space for all geometric operations.

Of the 184 subjects in the dataset, 179 were processed successfully and 5 were excluded for missing input files. The three benchmark subjects (100610, 102311, 102816) were taken in ascending identifier order from those 179 to give a computationally tractable benchmark, without reference to any anatomical criterion, so the six subject-hemisphere units are not selected for how well any method performs on them.

B MODEL ARCHITECTURE AND HYPERPARAMETERS

Differentiable reimplementation of $\mathcal{S}$. The three changes of Section 4.1, in detail. (i) Surface distances, originally obtained by ray-mesh intersection, are precomputed on a grid over the (α, β) search space and recovered at arbitrary (α, β) by differentiable bilinear interpolation. We did not quantify the interpolation error separately; the aggregate disagreement between the two implementations is reported in Appendix D. (ii) Hard nearest-voxel assignment of receptive field properties to each contact is replaced by a Gaussian-weighted average over V1 voxels. With x_i the position of contact i, v_j the V1 grey-matter voxels in the anatomical volume and $w_{ij} = \exp(-\|x_i - v_j\|^2 / 2\sigma_v^2)$ with $\sigma_v = 1.5$ voxels,

$$c_i = \min\Big(1, \textstyle\sum_j w_{ij}\Big), \qquad (\psi_i, \varepsilon_i, \rho_i) = \frac{\sum_j w_{ij}\,(\psi_j, \varepsilon_j, \rho_j)}{\sum_j w_{ij}},$$

where $c_i \in [0, 1]$ is the continuous validity score and $(\psi, \varepsilon, \rho)$ are polar angle, eccentricity and pRF size. Each hemisphere represents one contralateral hemifield, so polar angle within a unit spans half its range and the linear average does not cross the angular wrap-around. (iii) The binary activation mask used at inference is relaxed to $p_i = \mathrm{sigmoid}(\ell_i)$ of the electrode logits during training. The map is a linear superposition,

$$\tilde{P}(u) = \textstyle\sum_i p_i\, c_i\, \exp\big(-\|u - \mu_i\|^2 / s_i^2\big), \qquad P = \tilde{P} / \max_u \tilde{P},$$

with centre μ_i set by (ψ_i, ε_i) and width s_i by the cortical magnification factor at ε_i and the scale factor s (van der Grinten et al., 2024); at inference $p_i \in \{0, 1\}$. Yield is $Y = \frac{1}{N}\sum_i c_i$. The original implementation is NumPy based and relies on discrete voxel lookup and non-differentiable ray-surface intersection, which is what precludes gradient-based optimisation of ϑ in that pipeline.

Encoder. The target map (128×128, single channel) is processed by three convolutional downsampling stages followed by four residual blocks and a fully connected projection to a 256-dimensional latent vector, following de Ruyter van Steveninck et al. (2022) adapted to this input resolution. The electrode head is three fully connected layers mapping the latent vector to $N = 1000$ logits.

Placement parameter. $\vartheta \in \mathbb{R}^4$ is a single learnable parameter shared across all training examples within a subject hemisphere, initialised at the midpoint of its valid range (Table 2) and mapped into that range by a sigmoid.

Optimiser and schedule. All three training regimes of Section 4.3 use AdamW, batch size 8, up to 50 epochs with early stopping (minimum 10 epochs, patience 8 epochs on validation loss, minimum improvement 10^{-4}), a base learning rate of 10^{-3} for the encoder and 10^{-2} for the shared placement parameter ϑ, a learning-rate scheduler with patience 3 and reduction factor 0.5, and gradient-norm clipping at 1.0. Random seed 42 is used for weight initialisation, dataset splitting, and stochastic components throughout. Each of the three regimes trains for exactly 50 epochs at this configuration (Section 6.4); no per-regime hyperparameter re-tuning was performed.

Bayesian search (BO, Overlap). Placement search runs Gaussian process optimisation for up to 150 evaluations per target with early stopping on convergence, activating all V1-intersecting contacts. Overlap starts from the placement found by this search and additionally activates any V1-intersecting contact whose rendered phosphene centroid falls on target intensity above 4×10^{-4}, applied uniformly across all benchmark targets.

C BENCHMARK TARGET CONSTRUCTION

Two target domains are constructed at 128×128 resolution under a circular right-hemifield mask. The letter-like domain is derived from EMNIST Balanced (Cohen et al., 2017) (47 classes) passed through the pretrained encoder of van der Grinten et al. (2024) to obtain physically realisable phosphene maps, spatially jittered by a random vertical offset (seed 42). The geometric domain samples annular sector maps with four independent parameters: start angle and end angle within $[-90°, 90°]$ (right hemifield) with a minimum angular span of $45°$, and inner and outer radius within $[0, 64]$ px with a minimum radial span of 32 px, composed with a Gaussian intensity envelope and the same right-hemifield mask. Four fixed geometric targets reproduce the canonical arc stimuli of van Hoof et al. (2025) (upper sector, lower sector, inner ring, and complete Gaussian coverage) and are held out of the sampled pool.

Table 2: Implant placement search ranges for the placement vector ϑ of Section 3. Ranges follow van Hoof et al. (2025); the electrode activation mask is introduced in this work.

Parameter	Description	Range	Unit
α	Lateral insertion angle	$[-90, 90]$	degrees
β	Vertical insertion angle	$[-15, 110]$	degrees
Offset	Cortical surface to first contact	$[0, 40]$	mm
Shank length	Length of electrode shank	$[10, 40]$	mm
m_i	Per-electrode activation mask	$\{0, 1\}$	–

The β range shown is the one the learned placement uses in both hemispheres. The search-based baselines, run in the reference implementation, use it for the left hemisphere and $[-110, 15]$ degrees for the right.

Training and validation corpora are generated independently per domain with disjoint random seeds from the evaluation pool (train seed 0, evaluation seed 999 for the geometric domain); no evaluation target appears in any training or validation corpus. Each domain-specific corpus contains 112,800 generated maps, split 80/10/10 into training, validation, and an internal held-out test split at training time (90,240/11,280/11,280), matching the corpus size used for PhosInv-mixed's balanced 56,400-per-domain construction (Section 4.3). The evaluation pool comprises 1,410 letter-like targets (EMNIST Balanced test split, 47 classes $\times$ 30, class-stratified) and 1,410 geometric targets (1,406 sampled targets plus the 4 fixed arc targets described above), for 2,820 targets total across the two domains.

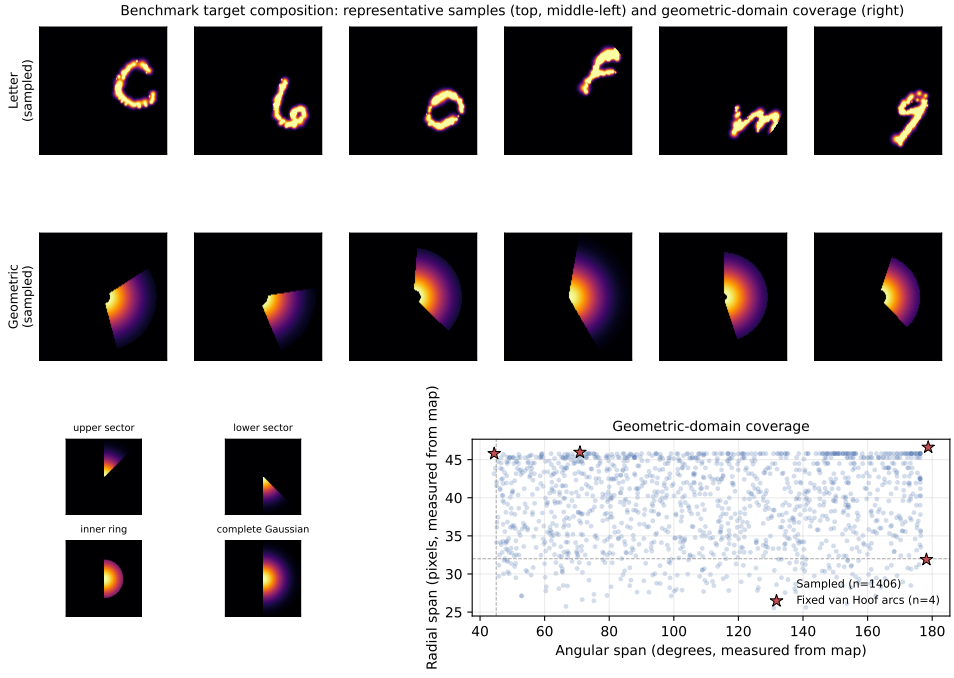

Figure 2: Benchmark target composition. Top and upper-middle rows: six representative sampled targets from each domain. Lower-middle left: the four fixed geometric targets reproducing the canonical arc stimuli of van Hoof et al. (2025). Right: angular versus radial span of the above-threshold pixel footprint of every geometric-domain target, measured directly from the rendered map, for the sampled pool (n=1,406) and the four fixed targets; dashed lines mark the sampling procedure's minimum angular (45°) and radial (32 px) span.

D ADDITIONAL RESULTS

Evidence for Section 6.1. Table 3 gives, for the full seven-contrast family of each domain, the quantities the main text treats as primary (mean difference, sign count and per-unit range over the six subject hemispheres) followed by the secondary p-values of Section 5.5.

Table 3: The seven-contrast family of each domain. Δ: mean over the six subject-hemisphere units of the per-unit difference in mean loss (negative favours the first method); Sign: number of units in which the difference is negative; Range: smallest and largest per-unit difference. These three are primary. Secondary, directional only: W: cluster-level Wilcoxon signed-rank p, Holm-corrected within the domain (its floor is $0.031 \times 7 = 0.219$); R: cluster-robust OLS p (loss $\sim$ method, standard errors clustered on the six units), uncorrected and Holm-corrected. BO and Overlap are per-target unless marked fixed-ref.

Domain	Contrast	Δ	Sign	Range	W (Holm)	R	R (Holm)
Geo.	PhosInv-letter vs. BO	-0.108	4/6	$[-0.300, 0.197]$	0.313	0.223	0.223
Geo.	PhosInv-geometric vs. BO	-0.284	6/6	$[-0.491, -0.173]$	0.219	< 0.001	< 0.001
Geo.	PhosInv-mixed vs. BO	-0.224	5/6	$[-0.369, 0.047]$	0.281	0.004	0.015
Geo.	PhosInv-letter vs. Overlap	-0.422	6/6	$[-0.553, -0.134]$	0.219	< 0.001	< 0.001
Geo.	PhosInv-geometric vs. Overlap	-0.598	6/6	$[-0.811, -0.473]$	0.219	< 0.001	< 0.001
Geo.	BO vs. BO (fixed-ref.)	-0.108	4/6	$[-0.427, 0.013]$	0.313	0.111	0.222
Geo.	Overlap vs. Overlap (fixed-ref.)	-0.090	5/6	$[-0.310, 0.010]$	0.250	0.056	0.168
Letter	PhosInv-letter vs. BO	-0.426	6/6	$[-0.632, -0.250]$	0.219	< 0.001	< 0.001
Letter	PhosInv-geometric vs. BO	-0.056	4/6	$[-0.236, 0.088]$	0.313	0.223	0.223
Letter	PhosInv-mixed vs. BO	-0.380	6/6	$[-0.739, -0.206]$	0.219	< 0.001	< 0.001
Letter	PhosInv-letter vs. Overlap	-0.532	6/6	$[-0.764, -0.304]$	0.219	< 0.001	< 0.001
Letter	PhosInv-geometric vs. Overlap	-0.162	6/6	$[-0.368, -0.007]$	0.219	0.005	0.009
Letter	BO vs. BO (fixed-ref.)	-0.227	6/6	$[-0.403, -0.051]$	0.219	< 0.001	< 0.001
Letter	Overlap vs. Overlap (fixed-ref.)	-0.217	6/6	$[-0.376, -0.085]$	0.219	< 0.001	< 0.001

Per-domain metric breakdown. Table 1 in the main text gives the complete per-domain loss, Dice, and active-ratio breakdown for all eleven methods; Hellinger distance and yield, omitted there for space, are available in the released benchmark output alongside every reported figure.

Fixed-reference sensitivity. Reference targets for the fixed-reference condition are drawn at the 10th, 50th, and 90th percentile of BO's own optimisation objective within each domain. Table 4 reports all six resulting variants, rendered under the reference NumPy simulator (Section 4.1) rather than the unified pipeline of Section 5.4, since only the 50th-percentile variant used in the main text was re-rendered under the differentiable simulator; the absolute values in this table are therefore not directly comparable to Table 1 and should be read only for the relative spread across percentiles. Loss ranges from 2.031 to 2.310 across all six variants and both domains, and the 50th-percentile variant is not a neutral median in the letter domain, where it is in fact the strongest (lowest-loss) of the three percentiles for both BO and Overlap.

Table 4: Fixed-reference sensitivity to which target's placement is used as the frozen reference, by percentile of BO's own search objective. Rendered under the reference NumPy simulator (see caveat above).

Method	Domain	p10	p50	p90	Range
BO (fixed-ref.)	Geometric	2.080	2.128	2.207	0.127
BO (fixed-ref.)	Letter	2.219	2.185	2.310	0.125
Overlap (fixed-ref.)	Geometric	2.031	2.148	2.225	0.194
Overlap (fixed-ref.)	Letter	2.149	2.126	2.198	0.073

Cross-renderer agreement. Section 6.5 summarises agreement between the two simulator implementations on 6,912 matched configurations. Under the NumPy metric convention, rank correlation is 0.546 for Dice and 0.611 for composite loss, with mean absolute differences of 0.130 (Dice) and

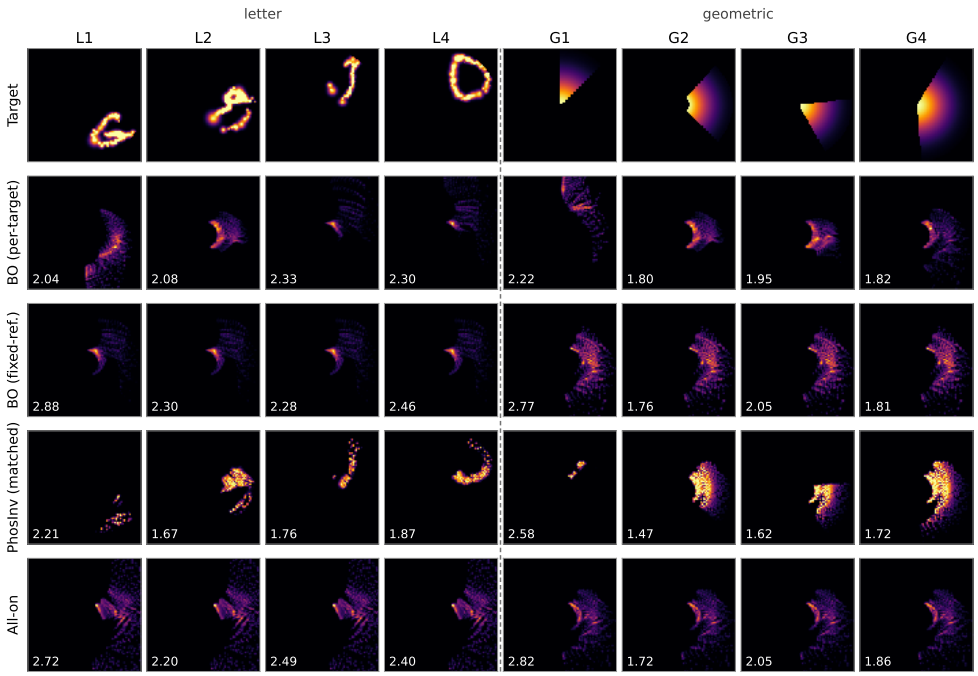

Figure 3: Reconstructions for eight benchmark targets in one subject hemisphere (102311_RH), rendered through the single pipeline of Section 5.4; the number inset in each map is its composite loss. Nothing here is hand-picked. The hemisphere is the one closest to the median of the six in *both* domains by matched PhosInv loss, being the letter median exactly and 0.002 off the geometric median; that rule reads only PhosInv's own loss and never the contrast against BO shown here. Target G1 is a canonical arc from van Hoof et al. (2025), kept fixed for comparability with that framework; the other seven are drawn from the evaluation sets with a fixed seed. No selection uses any method's loss, which is the circularity retired in Section 5.2. BO attains the lower loss in two of the eight columns, L1 and the fixed arc G1, against per-target BO-win rates of 7.2% on letter and 6.5% on geometric targets in this hemisphere; the aggregate advantage in Table 1 is a difference in means over 1,410 targets per domain, not a claim of per-target dominance. The *BO (fixed-ref.)* and *All-on* rows are constant within each block, the first because one reconstruction is applied to every stimulus and the second because activating every V1-intersecting contact ignores the target altogether. Only the PhosInv row varies with the target, which is the selection lever of Section 6.3 in visual form.

0.201 (loss); under each implementation's native convention, Dice rank correlation falls to 0.301. Agreement is domain-dependent, with roughly twice the mean absolute loss discrepancy on letter targets (0.113) as on geometric targets (0.058). Yield is not comparable across implementations at all (means of 0.42 versus 0.86 on identical inputs) and is reported under a single definition throughout, per Section 5.4.

Resampling analysis. Section 6.5 reports that across 500 stratified redraws of one third of the letter evaluation set (470 of 1,410 targets, 47 classes $\times$ 10), the sign of the PhosInv-letter minus PhosInv-geometric contrast and the PhosInv-letter minus BO contrast was preserved in all 500 draws, with median effect sizes of -0.370 and -0.204 respectively (standard deviation across draws 0.004 for both), agreeing with the full-sample values to four decimal places. An independent check using simple (non-stratified) random resampling at $n = 300$ gave the same result (0 sign flips, standard deviation 0.004–0.005).

Per-subject-hemisphere results. Table 5 gives mean loss by subject hemisphere for the four methods central to Sections 6.1 and 6.2, so that the sign counts reported there can be read against the underlying per-unit values. It is also the source for the left-hemisphere pattern noted in Appendix E.

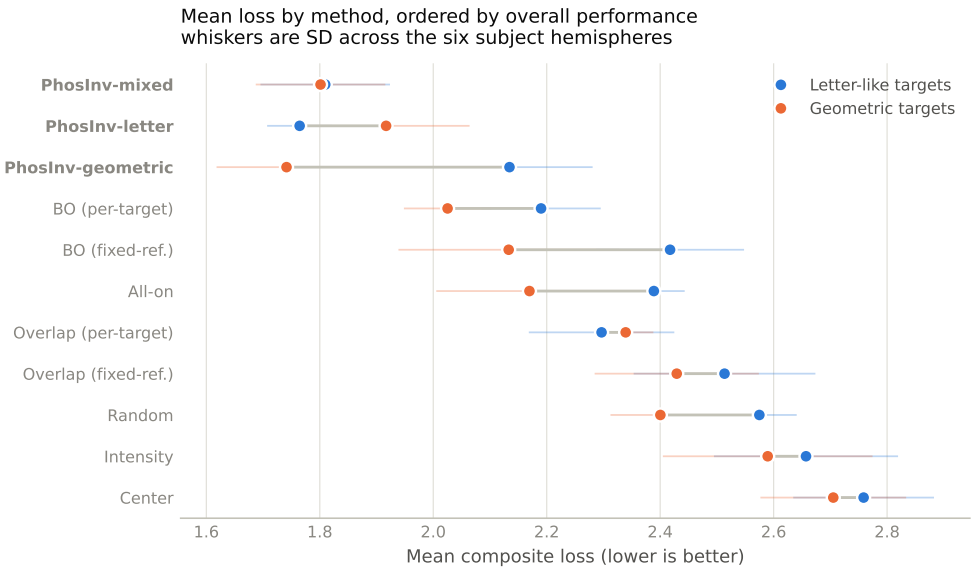

Figure 4: Mean composite loss by method, split by evaluation domain. Each row is one method, with a marker per domain joined by a rule whose length is that method's sensitivity to the evaluation domain; methods are ordered by their mean across the two domains, so the ranking can be read directly. Whiskers give one standard deviation across the six subject-hemisphere means and are drawn recessively, as the primary evidence is the paired per-unit comparison of Section 5.5 rather than this spread. The composite loss of equation 5 has no meaningful zero, so position rather than bar length encodes it here. Data underlying Table 1. Per-contrast statistical evidence for the comparisons discussed in this section is given in Appendix D.

Table 5: Mean composite loss by subject hemisphere, pooled across both evaluation domains, for the four methods discussed most extensively in the main text.

Method	100610_LH	100610_RH	102311_LH	102311_RH	102816_LH	102816_RH
BO (per-target)	2.187	2.194	2.137	1.991	2.135	2.003
PhosInv-letter	1.836	1.805	1.781	1.893	1.753	1.976
PhosInv-geometric	2.112	1.830	1.988	1.819	2.045	1.833
PhosInv-mixed	1.815	1.657	1.801	1.911	1.778	1.870

Holm-corrected pairwise contrasts. Table 3 lists all seven contrasts of each domain's family, with uncorrected and Holm-corrected values. Every contrast whose cluster-robust p is below 0.001 uncorrected stays below 0.05 after correction. The cluster Wilcoxon test cannot reach 0.05 after correction for any contrast, whatever the effect size, because of its 0.219 floor at this family size; it is reported only as a consistency check.

E Further Limitations

Left-hemisphere disadvantage. A left-hemisphere disadvantage appears in 7 of the 11 evaluated methods, including ones with no learned component, but is not consistent enough across methods to be called established (Table 5). We have no explanation for it and flag it as open.

What the simulator omits. The forward model omits temporal dynamics, cortical plasticity, and perceptual adaptation, and the array geometry ($10 \times 10 \times 10$ grid, 1 mm pitch) is fixed throughout. How the reported advantages would transfer to other geometries or channel densities is unknown, which matters particularly for the extrapolation to high-density arrays in Section 7.

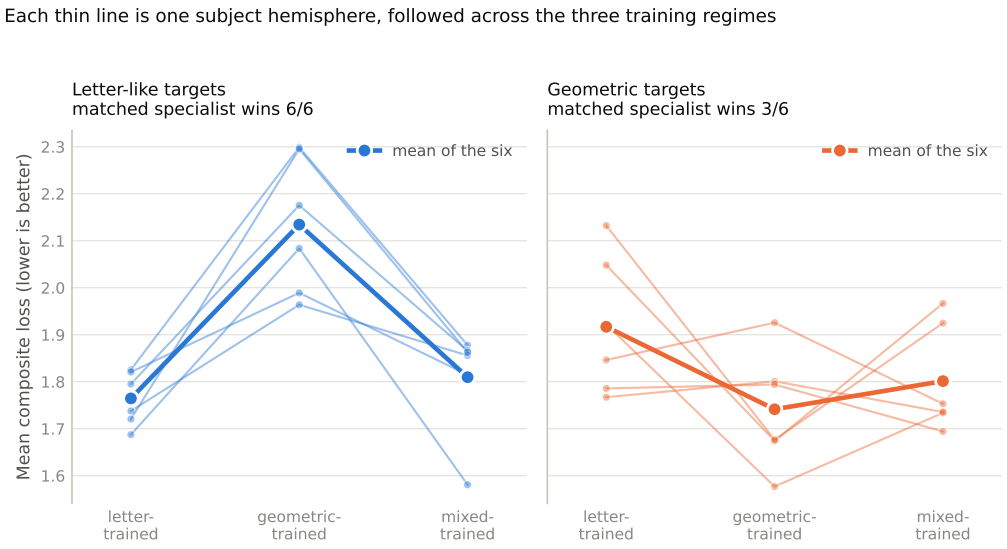

Figure 5: Cross-domain transfer by training regime. Each thin line follows one subject hemisphere across the three regimes, and the heavy line is the mean of the six; the six hemispheres are the same in both panels, so the comparison is paired and the lines can be read directly. On letter targets the six lines share a shape and the matched specialist wins in every hemisphere, whereas on geometric targets the lines cross and the matched specialist wins in only three of six, which is the asymmetry Table 6 reports and the reason the geometric-domain contrast is not read as an established effect. Panel headings quote the specialist-versus-specialist sign count; the separate comparison against PhosInv-mixed is given in the text.

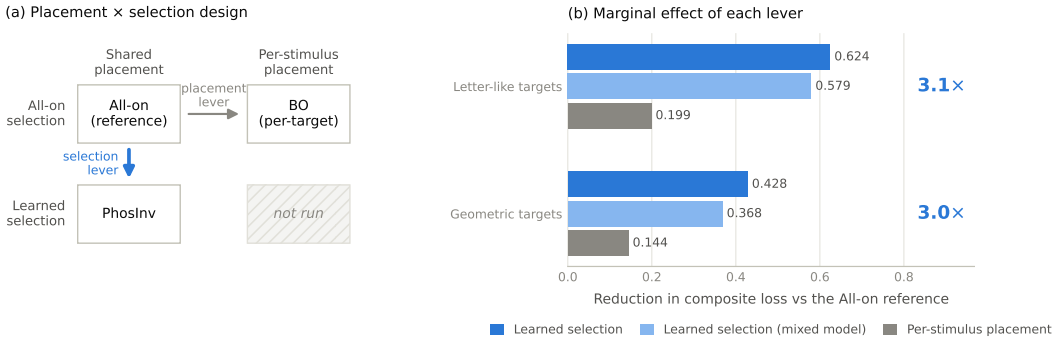

Figure 6: The two optimisation levers. **(a)** The placement-by-selection design of Section 3. Three cells are attainable; the fourth, per-stimulus placement combined with learned selection, would require a combinatorial selection search nested inside every placement optimisation and was not run, so the two effects are marginal rather than additive and no interaction term is estimated. Both levers are measured against the shared All-on reference, which holds architecture, training pipeline, and placement fixed. **(b)** Marginal effect of each lever, as a reduction in mean composite loss relative to that reference, computed at the level of the six subject hemispheres. Learned selection exceeds per-stimulus placement search by roughly a factor of three in both domains, and the ordering is unchanged when the mixed-corpus model supplies the selection cell instead of the domain-matched one.

Baselines not evaluated. The selection head is compared with search-based and non-learned selectors, not with stronger learned stimulus encoders such as hybrid neural autoencoders (Granley et al., 2022). Those methods learn stimulation for a fixed device geometry; the encoder here follows de Ruyter van Steveninck et al. (2022), and the contribution is the formulation of equation 3, which moves placement inside the learned problem, rather than the encoder design. A stronger encoder

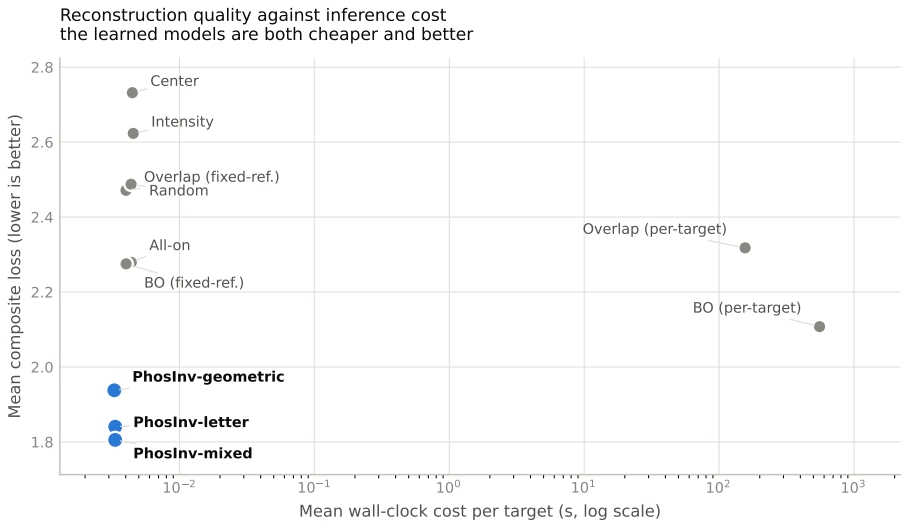

Figure 7: Reconstruction loss against mean per-target wall-clock cost (log scale), all eleven methods (mean over both domains). Three clusters are visible: per-stimulus search (BO, Overlap, hundreds of seconds), the fixed-reference variants (milliseconds, one-time reference cost already paid), and the learned or heuristic forward-pass methods (single-digit milliseconds). PhosInv variants combine the lowest loss with fast-cluster inference cost. The search-based methods are timed on a CPU with the NumPy simulator and the forward-pass methods on a GPU (Section 6.4), so the horizontal gap mixes amortisation with hardware and implementation differences and is not a matched speedup.

could replace f_ϕ in equation 4 unchanged, so the selection result is established against search and against non-learned selection, not against the best available learned encoder. Nor is the Overlap rule evaluated adaptively on a fixed placement: Overlap (fixed-ref.) freezes the mask as well as the placement, and only per-target Overlap, which refits its placement to every stimulus, applies the rule per stimulus.

Target domains. Evaluation covers letter-like and geometric targets only; natural scenes and clinical acuity optotypes such as Landolt C or tumbling E are not evaluated. Since a solution of equation 3 is defined relative to $\mathcal{D}$, and the transfer results of Section 6.2 show that coverage limits generalisation, such targets would need to enter the training distribution rather than be assumed to transfer.

Discretisation of the mask. Training uses the sigmoid relaxation of the mask and inference thresholds it at 0.5. No straight-through estimator or temperature annealing is used to align the two, and the resulting gap between soft and binarised Dice is documented in Appendix M; such techniques might narrow it but are not evaluated.

F YIELD CONVENTION

Yield is $Y = n_{\text{hits}}/n_{\text{contacts}}$, the proportion of contacts that intersect V1 grey matter and therefore evoke a phosphene, with $n_{\text{contacts}} = N = 1000$ throughout: the full array rather than the activated subset. We take n_{hits} from the continuous validity score of the differentiable pipeline rather than the reference implementation's hard voxel intersection. The two disagree on the hard/soft convention and, as a consequence, on the resulting mean (Appendix M), so this single definition is applied uniformly to every method reported in the paper.

G FIXED-REFERENCE TARGET SELECTION

Reference targets for the fixed-reference condition (Section 5.3) are drawn at the 10th, 50th and 90th percentile of BO's own optimisation objective within each domain, giving six variants. That

selection uses only the search method's own objective and is independent of PhosInv, so it cannot inherit the circularity noted in Section 5.2.

The main text reports the 50th-percentile variant in both domains, this being the only one re-rendered under the unified pipeline of Section 5.4. It is not uniformly the strongest of the three: under the reference renderer the 10th-percentile variant is lower-loss on geometric targets by 0.048, while the 50th-percentile variant is the strongest on letter targets. No single percentile rule is therefore uniformly conservative, and the baseline as reported is mildly favourable to us in the geometric domain. All six variants appear in Table 4.

H CROSS-DOMAIN TRANSFER TABLE

Table 6: Cross-domain transfer. Mean composite loss for each of the three training regimes (rows), evaluated on each domain (columns). Diagonal cells are domain-matched; off-diagonal cells evaluate a specialist outside its training domain. All entries share the placement-and-selection architecture of Section 4.2 and the single rendering pipeline of Section 5.4; only the training distribution differs across rows.

Training regime	Eval. on geometric	Eval. on letter
PhosInv-letter	1.917	1.764*
PhosInv-geometric	1.741*	2.134
PhosInv-mixed	1.801	1.810

*Domain-matched (diagonal) cell.

I LEVER DECOMPOSITION CAVEATS

Two caveats bound Section 6.3. First, the fourth cell of the placement-by-selection design, per-stimulus placement combined with learned selection, would require a combinatorial selection search nested inside every placement optimisation and was not run, so the two effects are marginal rather than additive and no interaction term is estimated. Second, the decomposition is reported per domain and not pooled across domains, because among the two specialist regimes there is no single shared-placement model that is domain-matched in both domains simultaneously. PhosInv-mixed relaxes this only partially: it is not domain-matched in the same sense (Section 6.2), so its selection-effect row is reported alongside the domain-matched row rather than merged with it. Third, the search-once row uses the 50th-percentile reference target of Appendix G; how the comparison moves with the reference is bounded only under the other renderer (Table 4), where the spread across percentiles (up to 0.127 for BO) is larger than either mean search-once effect, so the comparison is specific to this reference; a better-chosen reference could widen the searched placement's advantage. Per-unit differences (learned minus searched) are heterogeneous (geometric -0.504 to 0.312, searched placement better in 5/6; letter -0.336 to 0.125, 3/6).

J STATISTICAL ANALYSIS IN FULL

Each method yields one loss value per target and subject hemisphere, so observations are non-independent along two axes: repeated targets within a subject hemisphere share one implant placement (and, for the learned methods, one model), and the same 2,820 targets recur across methods. A fully crossed mixed-effects model with subject (nested hemisphere) and target as random effects was our intended primary analysis, but at this scale ($n = 8{,}460$ per domain, up to $1{,}410$ target levels) the standard restricted-maximum-likelihood fitting procedure did not converge reliably, returning between-unit variance estimates at the boundary of the parameter space and correspondingly untrustworthy standard errors. We verified this failure directly by cross-checking its p-values against the cluster-level test on contrasts where the two should agree, and found cases where the mixed-effects fit reported $p < 0.0001$ for a contrast whose direction was not even consistent across the six subject-hemisphere units. We do not report numbers from that fit.

Table 7: Decomposition of the two optimisation levers, reported separately per evaluation domain and additionally for PhosInv-mixed, which supplies a shared-placement selection cell in both domains simultaneously. Upper block gives mean composite loss for the attainable cells of the placement-by-selection design. Lower block gives the marginal effect of each lever relative to the common All-on reference, as a loss reduction, with larger values indicating a stronger lever; a negative value means the alternative placement is worse than the learned one. All entries are rendered through the single pipeline of Section 5.4.

	Letter	Geometric
All-on (shared placement, all-on selection)	2.389	2.170
PhosInv (shared placement, learned selection)	1.764	1.741
PhosInv-mixed (shared placement, learned selection)	1.810	1.801
BO (per-stimulus placement, all-on selection)	2.190	2.025
BO fixed-ref. (placement searched once, all-on selection)	2.418	2.133
Effect of learned selection (placement fixed), PhosInv	0.624	0.428
Effect of learned selection (placement fixed), PhosInv-mixed	0.579	0.368
Effect of per-stimulus placement (selection fixed)	0.199	0.144
Effect of search-once placement (selection fixed)	−0.029	0.037

The cluster-robust ordinary least squares used instead is stable at this design and gives standard errors an order of magnitude larger than an unclustered fit on the same data, consistent with the non-independence the clustering is meant to correct for. At six clusters such standard errors are nonetheless known to be liberal, and we apply no small-cluster correction (such as CR2 or a wild cluster bootstrap), so every cluster-robust p-value in this paper is exploratory and directional. The six clusters are also the two hemispheres of each of three subjects, which share subject-level anatomy, so they are not fully independent and the effective number of independent units may be closer to three. The companion cluster-level Wilcoxon signed-rank test has the opposite failing: its smallest attainable two-sided p at $n = 6$ is 0.03125, and after Holm correction within a seven-contrast family it cannot reach significance regardless of effect size. For these reasons neither test carries a claim. The main text rests each contrast on its effect size, its sign count over the six units and the range of the per-unit differences (Table 3); a contrast is read as established only when its sign holds in all or all but one of the six units. The fixed random seed governs dataset splitting, model initialisation, and every stochastic selection rule.

K Inference cost in full

Measurement conditions. The forward-pass figure quoted in Section 6.4 is 7.3 ms for PhosInv-geometric and 11.1 ms for PhosInv-letter, corresponding to 137 and 90 Hz for the model and simulator alone, measured on a single NVIDIA GeForce RTX 5070 over 200 timed single-sample iterations per model set on both hemispheres of one subject. Latency was not re-measured on the remaining two subjects, whose anatomies enter the loss results but not this timing. Per-stimulus Bayesian search requires 555.4 s under the same benchmark, measured under the reference NumPy implementation of $\mathcal{S}$ that BO's search is actually run against (Section 4.1), on a multi-core CPU host rather than the GPU used for the amortised model. The two therefore differ in hardware and in simulator implementation as well as in method, and the ratio between them is not a matched speedup: running the search on a GPU or through the PyTorch simulator would shorten it, and running the forward pass on a CPU would lengthen that. We did not measure either under matched conditions and do not report a matched figure.

Correction to a previously reported figure. An earlier latency of 81 ms for this pipeline was obtained by dividing a batched inference time by the batch size, which does not measure single-sample latency. The value reported here is measured directly on single samples and is faster, not slower, than the earlier figure.

Amortisation. Training solves equation 4 once per subject hemisphere at a mean cost of 72,768 s (20.2 h; range 19.2 to 21.1 h across the six units, on one RTX 5070 each). That figure covers the training run itself and does not include hyperparameter selection, which was fixed in advance from

the specialised-model configuration and not re-tuned per regime (Appendix B). The crossover at 131 stimuli is 72,768 s divided by 555.4 s per BO query; the numerator is GPU time and the denominator CPU time under the NumPy simulator, so the crossover inherits the same hardware and implementation difference and would move under matched conditions.

Active contacts. Counts are of contacts commanded on. BO and All-on command all $N = 1000$, of which only the fraction Y intersecting V1 evokes a phosphene, so against V1-intersecting contacts alone the reduction is smaller than against N. Averaged over both evaluation domains the per-model means are 119.5 contacts (PhosInv-geometric), 153.1 (PhosInv-letter) and 168.3 (PhosInv-mixed), of the $N = 1000$ available. The per-domain domain-matched values quoted in the main text are 170.5 on geometric and 124.6 on letter targets.

L ELECTRODE SELECTION ABLATION IN FULL

This appendix expands the selection ablation summarised at the end of Section 6.3. Five conditions share the placement learned by PhosInv and differ only in how the activation mask is chosen. All-on activates every V1-intersecting contact and so removes the selection head while leaving the architecture, the training pipeline, and the placement untouched. Random, Intensity, and Center each activate k contacts matched to the mean active count of PhosInv, so they hold the number of stimulated sites fixed and vary only the rule.

Table 8: Electrode selection ablation. All five conditions share the placement learned by the domain-matched PhosInv model (PhosInv-geometric for the geometric column, PhosInv-letter for the letter column). Random, Intensity, and Center activate k contacts matched to the mean active count of the domain-matched PhosInv model in that domain ($k = 171$ geometric, $k = 125$ letter).

	Geometric		Letter	
Condition	Loss	Dice	Loss	Dice
PhosInv	1.741	0.642	1.764	0.655
All-on	2.170	0.395	2.389	0.303
Random	2.400	0.256	2.575	0.190
Intensity	2.590	0.146	2.657	0.134
Center	2.705	0.090	2.759	0.075

Each heuristic fails in a characteristic way, with random selection producing spatially incoherent scatter, intensity-based selection over-committing to the brightest region of the target, and centre-based selection ignoring the target entirely. This isolates the learned mapping from target structure to electrode subset as the operative mechanism, and does so within a single architecture and training pipeline rather than across method families.

The three heuristics are simple by construction, so they bound the value of selection from below rather than establishing what a strong non-learned selector could achieve. Overlap supplies that upper comparison in Section 6.1, selecting by explicit spatial correspondence between rendered phosphene locations and the target rather than by a naïve rule: learned selection still leads, by a wider margin than against All-on, in both domains and every subject hemisphere. The dominant-lever conclusion therefore does not rest on the weakness of the four baselines considered here.

M ROBUSTNESS CHECKS IN FULL

This appendix gives the four checks summarised in Section 6.5 at full length. Nothing here was dropped in compression; it was relocated because none of it changes a reported contrast, only bounds it.

Size of the evaluation sample. Drawing 500 independent stratified subsamples of one third of the letter evaluation set, the sign of the two headline PhosInv contrasts was preserved in 500 of 500 draws, with median effect sizes matching the full-sample values to four decimal places (Table 3 and

the resampling paragraph of Appendix D); the reported contrasts are not an artefact of which targets were sampled.

Choice of fixed reference. Across all six fixed-reference variants of Section 5.3, mean loss ranges from 2.031 to 2.310 under the reference renderer (Table 4); every variant, including the best of the six, is worse than the domain-matched inverse model in the corresponding domain by a margin larger than the 0.194 total spread across percentiles, so the conclusion does not rest on the particular reference chosen. This is the one comparison in the paper that spans two renderers, and it is reported only because the spread it measures is small relative to the contrast it bounds. Notably, the strongest reference is not at the same percentile in both domains, and within the letter domain the 10th-percentile reference performs worse than the 50th: how well a target is reconstructed under its own optimisation is a poor predictor of how well its frozen configuration transfers to others, a direct consequence of equation 2 and equation 3 not being the same objective.

Agreement between simulator implementations. Rendering 6,912 matched configurations through both implementations of $\mathcal{S}$ with implant parameters, activation, and target held fixed, the two agree only moderately at the level of individual configurations (Dice rank correlation 0.546, composite-loss rank correlation 0.611; full breakdown, including the domain-dependence of the discrepancy, in Appendix D). This magnitude is why Section 5.4 requires every method to be scored under a single implementation: scoring different method families under different implementations would confound electrode configuration with renderer at a scale comparable to the effects being measured. It also means the benchmark's conclusions are specific to the renderer they are scored under. The search-based methods were fitted to the reference implementation and PhosInv to the differentiable one, and at this level of agreement a method is expected to do better under the implementation it was fitted to; we therefore make every claim under the differentiable renderer and do not claim that the ranking carries over to the reference implementation, which we have not re-scored in full.

Yield is not comparable across implementations for identical inputs (means of 0.42 and 0.86) and is therefore reported under the single definition of Section 5.4 throughout, which depends only on placement and not on the activation mask. An earlier claim, made for the 30-target benchmark under a mismatched pair of conventions, that yield accounts for roughly 30 percent of the PhosInv minus BO gap does not carry over: recomputed under the single definition used here, yield contributes 2.1 to 3.2 percent of that gap across domains, and is retracted rather than restated.

Soft versus binarised Dice. The training and evaluation loss formulations of Section 4.2 can rank the same two conditions differently. For PhosInv-geometric, soft Dice favours the left hemisphere while binarised Dice favours the right, an outright reversal, whereas for PhosInv-letter the two agree in direction. Recomputing the training-time metric in its binarised form reproduces the benchmark values to within 0.013, confirming training and evaluation were never inconsistent and that the apparent conflict is entirely a property of comparing two different formulas rather than a defect; checkpoint misassignment and any influence of the fixed arc targets were separately excluded. All results in this paper use the binarised form throughout.

N PROSPECTIVE CALIBRATION OF THE PHOSPHENE SIZE SCALE FACTOR

This calibration is exploratory and is not applied to any result in this paper; every experiment reported here runs the simulator at the default $s = 0.5$ of van der Grinten et al. (2024), which was never tuned against reconstruction loss. We separately explored whether s could instead be estimated from first principles, by coupling a static finite-element solution for the extracellular field around a single micro-ring electrode to compartmental simulation of a layer-5 pyramidal neuron (Mainen & Sejnowski, 1996) in NEURON (Carnevale & Hines, 2006), over a $1\,\mathrm{mm}^3$ lattice around the electrode. Under the two stimulation configurations this array geometry permits within one cortical voxel, the resulting activation radii were 0.094–$0.115\,\mathrm{mm}$, giving a biophysically derived scale $s_{\mathrm{bio}} = 0.245$–0.299 – substantially below the literature default, suggesting $s = 0.5$ overestimates the spatial extent of neural activation for this electrode geometry.

A comparison of $s = 0.5$ against s_{bio}, run on a scale-configurable simulator variant not used elsewhere in this paper, found lower composite loss under s_{bio} (paired Wilcoxon, $p < 0.05$). We do not lean on this: it was run on the 30-target benchmark that Section 5.2 retires for its circular diffi-

culty definition, so it inherits that defect and is evidence only that the two settings differ, not of how much. We report this calibration as a candidate refinement to be re-evaluated on the 2,820-target benchmark in future work, not as an established result of the present study.

The scale factor also bears on the selection result. At $s = 0.5$ phosphenes from neighbouring contacts at 1 mm pitch overlap substantially, which penalises dense activation and so may enlarge the margin of sparse learned selection over all-on activation (Section 6.3); at a smaller s phosphenes are more focal, and sparse selection could leave gaps that denser activation would fill. The direction and size of this effect on the selection-versus-placement ratio are untested. The compartmental model also covers only direct somatodendritic activation of one pyramidal morphology and omits passing axons, which typically have lower thresholds and broader recruitment. Full simulation parameters (lattice geometry, waveform, coupling equations) are provided in the code archive.