Journal of
Eye Movement Research

MDPI

Review

One Model or Many? Domain Transfer and Specialization in Predicting Gaze on Designed Media

Thomas Zoëga Ramsøy [1,2,3]

[1] Neurons Inc., 2630 Taastrup, Denmark; thomas@neuronsinc.com
[2] International Center for Applied Neuroscience, 4581 Rørvig, Denmark
[3] Free University of Languages and Communication (IULM), 20143 Milan, Italy

Abstract

Deep-learning models that predict aggregate fixation density from a media asset alone are now used as measurement instruments in advertising, packaging, interface, and educational research. The central validity question for this practice is domain transfer: whether a model trained on one distribution of stimuli, observers, tasks, and viewing contexts predicts fixations on the materials it is actually deployed on. This critical narrative review organizes the transfer problem into three progressively finer scales of stimulus shift—from natural scenes to designed media, between classes of designed media, and between categories within a class—and a cross-cutting axis of observers, tasks, contexts, and time, and finds the evidence sharply asymmetric: degradation from natural scenes to designed media is robustly documented; transfer between designed-media classes is supported by limited but consistent evidence; within-category and population-level transfer remain largely untested. I then reframe the practical response, specialization, as a finite-data pooling problem—complete pooling (one generalist), no pooling (independent specialists), and partial pooling (shared structure with category-specific adaptation)—and argue that partial pooling is the hypothesis most consistent with the evidence, not an established result. A company practice vignette, drawn from an unpublished internal development program, illustrates how the trade-off resolves operationally. I close with a pre-deployment transfer-testing protocol with predefined acceptance criteria and reporting items for studies that rely on predicted rather than measured gaze.

Keywords: gaze prediction; predictive gaze modeling; visual attention; saliency models; domain transfer; generalization; consumer neuroscience; neuromarketing; AI eye tracking; model validation

Check for updates

Academic Editor: Peter König

Received: 8 October 2026
Revised: 11 November 2026
Accepted: 17 November 2026
Published: 24 November 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

1.1. Predicted Gaze as a Measurement Instrument

Computational models of visual salience have accompanied eye-movement research for a quarter of a century [1,2], but their role has changed. For most of that period they were explanatory objects: hypotheses about the image-driven component of fixation selection, tested against recorded eye movements. With deep-learning models trained on large fixation datasets [3–5], prediction quality on benchmark natural-scene datasets rose substantially [6,7], and the capability was productized. Several commercial platforms now return predicted attention maps—and, in some cases, derived scores carrying construct labels—for any uploaded advertisement, package, webpage, or video, with no participant ever viewing the asset. Applied studies in marketing, education, and political communication increasingly treat these predictions as data [8–10].

Whether this substitution of predicted for measured gaze is sound depends on one question above all others: does a model trained on one distribution of stimuli, observers, and viewing conditions predict fixations on the materials it is actually deployed on? This is the domain-transfer question, and it is the subject of this review. It is distinct from how well models perform on the benchmarks they are developed against, and it is the question that matters for readers of applied studies, because the deployment domain—this year's advertisements, this client's packaging category, this market's audience—is essentially never the training domain.

1.2. What These Models Predict, and What They Do Not

Precision about the predicted quantity is a precondition for the rest of the argument, because the literature moves freely among "saliency," "visual attention," "predicted gaze," and "attention maps", which are not synonyms. The models at issue take an asset as input and output a spatial map that is, in the well-specified cases, an estimate of aggregate fixation density over a population of observers under free viewing—an estimate of where fixations from many viewers are likely to fall, integrated over an exposure [5,7]. Some systems additionally output derived quantities: percentages of predicted density falling within areas of interest, or scores with construct labels. What asset-only models do not recover—because nothing in the input supports it—is any individual observer's gaze, any task-specific scanpath, any temporal order of fixations in a static image, any pupillary or physiological response, and any psychological construct such as comprehension or cognitive demand, unless a separate validation links the predicted map to that construct.

The commercial term "predictive eye tracking" can suggest that eye movements are being tracked; they are not. Throughout, I use *gaze prediction* or *fixation-density prediction* for the model output, and reserve *attention* for the psychological construct, which predicted fixations index only under assumptions that must be stated [11].

The review also excludes gaze-estimation systems, such as webcam-based eye tracking, that measure the gaze of an actual, present viewer [12,13]. Those are measurement instruments, not predictive ones; they enter here only as possible sources of ground truth, at which point their own data quality becomes a reporting item (Section 6).

1.3. A Formal Statement of the Transfer Problem

It is useful to write the prediction setting explicitly. Let Y denote the fixation-density map, X the asset, T the task under which viewers observed it, O the observer characteristics, and C the viewing context—device, viewing geometry, exposure duration, surrounding content. Write $Z = (T, O, C)$ for the conditions collectively, and P_{train} for the joint distribution of assets and conditions from which the training corpus was drawn. A gaze-prediction model fitted to stimulus–fixation pairs estimates

$$f^*(x) = \int \mathbb{E}[\, Y \mid X = x,\ Z = z \,]\, dP_{\text{train}}(z \mid x), \tag{1}$$

that is, not $\mathbb{E}[Y \mid X, Z]$ but its average over the conditions that happened to hold in the training corpus. The free-viewing task, desktop presentation, and narrow participant demographics of benchmark ground truth are therefore part of the estimand, not a caveat about the data. Deployment substitutes a new joint distribution, and Equation (1) makes the three ways transfer can fail distinguishable: (i) the asset distribution $P(X)$ changes—a new stimulus class or category; (ii) the conditional distribution of conditions $P(Z \mid X)$ changes—a new task, audience, device, or exposure; or (iii) the mapping $\mathbb{E}[Y \mid X, Z]$ itself changes—as when a static model is applied to video, whose fixation dynamics no static mapping can represent. These are different failures with different remedies, and collapsing all three into covariate shift in X is a category error that the structure of Section 2 is designed

to avoid: Sections 2.1–2.3 treat shifts in $P(X)$; Section 2.4 treats shifts in $P(Z \mid X)$; and the change of mapping arises for video, in Sections 2.2, 3 and 5.2.

1.4. Scope, Review Approach, and Contributions

Scope. The review concerns designed media—advertisements, packaging, brand materials, interfaces, editorial and educational layouts—and treats the natural-scene benchmarking literature as background rather than subject matter [5–7].

Review approach. This is a critical narrative review, not a systematic one. Literature was identified in September 2026 through targeted keyword searches (Google Scholar and general web search) combining gaze-prediction terms ("saliency model", "saliency prediction", "fixation prediction", "visual importance", "attention prediction") with designed-media terms ("advertisement", "user interface", "webpage", "graphic design", "packaging", "video"), followed by backward and forward citation chaining from the included studies and from a recent review of fixation prediction [7]. Studies were assigned to a scale by the source and target domains of their evaluation. A study was treated as direct evidence for a scale if it evaluated a gaze- or saliency-prediction model's output against independently measured eye tracking on designed media, or on a stated subset of it (a device, format, or population); studies that validated models only against natural-scene benchmarks were treated as background (Section 1.1) rather than as transfer evidence for a given scale, and this criterion, not an independent systematic count, is what underlies every "sparse" or "largely absent" judgment below. Commercial validation reports were not sought systematically; the one included (Section 5) is my own company's and is reported as such. Because no formal screening protocol was applied, statements that evidence at a scale is "sparse" or "largely absent" mean that these searches identified none or few directly relevant studies, not that a systematic search established their absence.

Contributions. The review (i) organizes the transfer problem into three scales and a cross-cutting axis, with the formalization of Equation (1), and synthesizes the evidence at each (Section 2); (ii) separates established mechanisms of transfer failure in designed media from plausible but untested ones (Section 3); (iii) reframes the generalist-versus-specialist decision as a pooling problem, specifies the experiment that would test it, and offers a hypothesis-driven decision guide (Section 4); (iv) reports a company practice vignette illustrating the trade-off (Section 5); and (v) provides a pre-deployment transfer-testing protocol with predefined acceptance criteria and reporting items (Section 6).

1.5. Disclosure

I founded and lead Neurons, a commercial vendor of gaze-prediction tools, and Section 5 reports material from my company's internal development program. The containment measures are stated where they arise and in Section 7.4.

2. Scales of the Transfer Problem

The transfer question is usually posed once—"do saliency models work on ads?"—but the evidence is better organized by scale, because the answer, and the available evidence, differ (Figure 1; Table 1). Scales 1–3 are progressively finer shifts in the asset distribution $P(X)$; the fourth axis—observers, tasks, contexts, and time—is orthogonal to them and cuts across all three.

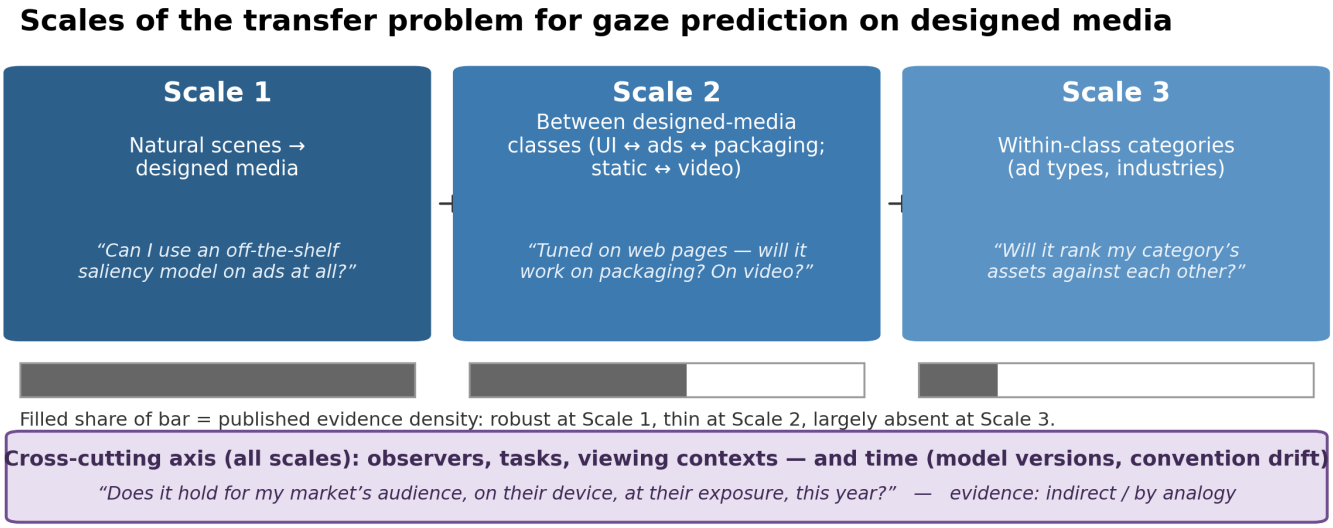

Figure 1. Scales of the transfer problem for gaze prediction on designed media. Three progressively finer shifts in the asset distribution (Scales 1–3) and a cross-cutting axis of observer, task, context, and temporal shift. Bars indicate the density of published evidence at each scale.

2.1. Scale 1: Natural Scenes to Designed Media

The foundational scale is the transfer from the photographic natural scenes that dominate fixation datasets [4,14] to designed artifacts. Here the evidence is consistent and the degradation substantial. Bylinskii and colleagues found that natural-image saliency models fit importance judgments on graphic designs and visualizations poorly, and that design-trained models substantially outperformed them [15]; follow-up work across design types reached the same conclusion [16]. For mobile user interfaces, a controlled study with 30 participants and 193 interfaces found that classic, parameter-free saliency models fit the data weakly, and that data-driven models improved significantly when trained on interface-specific data, with NSS rising from 0.66 to 0.84; the same study documented a strong bias toward the top-left of the display, toward text, and toward images, with bottom-up features such as color and size mattering less—a pattern the authors attribute to viewer expectations [17]. The UEyes study extended the point across interface types with 62 participants and 1980 screenshots, documenting systematic differences in fixation biases between webpages, mobile and desktop interfaces, and posters, and the benefit of type-appropriate training data [18].

2.2. Scale 2: Between Classes of Designed Media

The second scale asks whether a model tuned for one class of designed media transfers to another: interfaces to advertisements, static pages to video. Direct evidence is limited but consistent in direction. UEyes shows that even within "interfaces," fixation-bias structure differs by type [18]. Video introduces a qualitatively different problem: gaze in dynamic scenes exhibits attentional synchrony across viewers, with initial saccades defaulting toward the screen center, regions of high motion, and people before task-relevant features take over [19], and gaze clustering across viewers is predicted above all by motion and flicker [20]. Static models applied frame by frame represent none of this. This is a change in the conditional mapping of Equation (1), not only in $P(X)$. The vignette in Section 5 offers convergent grey-literature evidence.

2.3. Scale 3: Within-Class Categories

The third scale is the least studied and the most consequential for applied use: do categories within a class—print versus banner advertising; beauty versus automotive versus grocery—differ enough that a class-level model mispredicts or, more subtly, misorders assets within a category? Two considerations make within-class heterogeneity plausible a priori. Categories differ systematically in the features that drive designed-media fixations (text density, face presence, brand-element placement, clutter); consistent with this, the internal program in Section 5 found that an entropy-based image-complexity index differed significantly across verticals [21]. And range restriction operates within categories: a model may rank a heterogeneous set of assets well while failing to discriminate among the professionally produced, convention-following assets of one category—which is exactly the comparison applied users care about [22]. The searches described in Section 1.4 identified no published study that tests within-category discrimination of gaze-prediction models against measured eye tracking on advertising material; the vignette in Section 5 speaks to this scale and reports an equivocal picture.

Table 1. The transfer scales as an axis matrix.

Axis	Source Domain	Target Domain	Shift Type (Section 1.3)	Evidence State
Scale 1	Natural-scene photographs	Designed media (designs, interfaces, ads)	$P(X)$	Robust: substantial degradation; domain training helps [15–18]
Scale 2	One designed-media class; static images	Another class; video	$P(X)$; mapping (for video)	Limited but consistent [18–20]; grey-literature convergence (Section 5)
Scale 3	Class-level model	Categories within the class	$P(X)$, with range restriction	Largely untested in the published record; vignette equivocal (Section 5)
Cross-cutting (a)	Free viewing, desktop, narrow demographic	Applied task, device, audience	$P(Z \mid X)$	Indirect; motivation by analogy [23–28]
Cross-cutting (b)	Model version at validation	Model version at use	Governance	Documented in general machine learning [29,30]; unstudied for gaze prediction

2.4. The Cross-Cutting Axis: Observers, Tasks, Contexts, and Time

This axis is best split in two.

Observers, tasks, and contexts. Fixation behavior differs with task [23,24] and with cultural background in scene viewing [25]; benchmark ground truth is overwhelmingly free viewing, at multi-second exposures, on desktop displays, by observers from a narrow demographic range. For applied deployment, the audience, task (find the price; evaluate the offer), exposure (a feed scroll measured in hundreds of milliseconds), and device (a phone held at 30 cm) routinely differ from all of these. Direct evidence on subgroup disparities in gaze-prediction models was not identified by the searches described in Section 1.4; the concerns about WEIRD sampling [26] and demographically skewed training data in other machine-learning systems [27] are cited here as motivation by analogy, not as evidence that gaze-prediction systems exhibit the same disparities. The nearest direct evidence is on cross-dataset generalization: the strongest current natural-scene model was explicitly developed to give calibrated predictions out of domain, precisely because earlier models

did not [28], and the benchmarking literature treats dataset-specific bias as a first-order problem [7].

Time. Media conventions drift, and commercial models are updated continuously and often silently, so the system evaluated last year is not the system used today [29,30]. This is a model-governance problem rather than a transfer problem in the statistical sense, but it has the same consequence for the applied user: a validation result has a date.

3. Why Transfer Fails: Mechanisms and Hypotheses

Understanding *why* transfer degrades indicates when a failure should be expected and what a fix requires. This section separates mechanisms with direct evidence from those that are plausible but, for gaze prediction on designed media specifically, untested; the latter are marked as hypotheses. The mechanisms are made visible by comparing what models of different lineages characteristically output for the same stimuli (Figure 2): a biologically inspired feature-contrast model [1], a graph-based low-level model [31], a deep model trained on natural-scene fixations [3,7], and a model trained on designed commercial media.

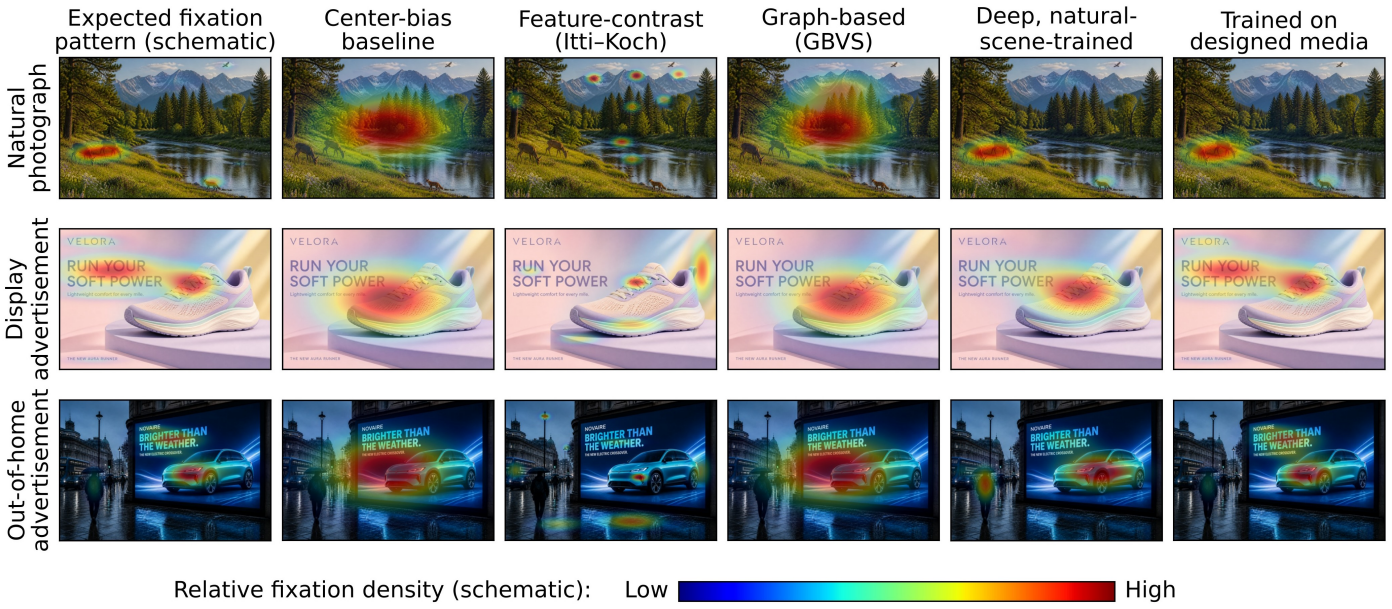

Figure 2. Schematic illustration of characteristic output patterns of four model lineages on three stimulus types. Rows: a natural photograph, a display advertisement, and an out-of-home advertisement in a street scene (AI-generated fictional-brand stimuli). Columns: the expected human fixation pattern, a center-bias baseline, and the characteristic outputs of a feature-contrast model (Itti–Koch), a graph-based model (GBVS), a deep natural-scene-trained model, and a model trained on designed media. Overlay colour shows relative fixation density on a common schematic scale from low (blue) to high (red); regions without overlay have negligible density. All maps are schematic renderings of the patterns predicted by the mechanisms discussed in this section; they are not model outputs or recorded data. Note that the center-bias baseline fails on the out-of-home row because the stimulus includes the wide street context surrounding the billboard: that context shifts the true region of interest away from the image's geometric center, so a baseline centered on the whole scene misses the billboard itself—the failure mode that makes baselines indispensable in transfer evaluation.

Text (evidence). Text attracts fixations robustly and early, over and above low-level salience: in free viewing, observers look at text regions many times more often than at size- and position-matched control regions, and text is nearly as effective as faces in this respect [32]; the bias is robust in real-world scenes [33]. Natural-scene training sets contain comparatively little text, so learned models under-weight it, and hand-crafted models never represented it. Whether the design-domain training gains at Scale 1 [15–17] are carried

specifically by learned text and layout features is a plausible hypothesis that would require ablation to establish.

Faces (evidence) and brand elements (hypothesis). Faces draw fixations across stimulus types and tasks [32]. Logos and product shots are also fixated in advertising, but the mechanism is contested: visual form, semantic familiarity, and goal relevance are distinct explanations, and attributing the effect to learned brand behavior would require evidence separating them. What is not contested is the training-data implication: a model can only learn category-typical brand-element placement from data containing that category.

Layout conventions and reading direction (evidence mixed). Genre-specific fixation-bias structure is documented across interface types [18], and the mobile-interface evidence points to viewer expectations as the source [17]. Reading-pattern conventions for webpages exist but are not universal laws; they depend on layout, language and reading direction, device, and task. Text legibility at the deployed size, not merely text presence, plausibly governs whether text is fixated (hypothesis).

Task, exposure, and context (evidence). Fixation behavior is task-dependent [23,24]. Benchmark ground truth is free viewing at multi-second exposures; applied viewing is often goal-directed and brief. A model can predict free-viewing fixation density well and still mispredict a goal-driven, sub-second exposure. Isolated-asset presentation (an advertisement alone on a screen) and in-context presentation (the same advertisement in a feed or on a street) are different conditions C with different priors on where viewers begin. Whether partial pooling can bridge this gap—conditioning a shared representation on task as an input, or fine-tuning a free-viewing-trained model on a small goal-directed sample (Section 4.2)—is itself an open instance of the pooling problem; nothing in the free-viewing literature establishes that the passive visual pull documented above (text, faces) and active task relevance combine in a way any such adaptation can recover, and the two should not be assumed to simply add.

Central-bias structure and viewing geometry (evidence). A large share of naive agreement on any stimulus class is carried by generic viewing biases, above all central fixation bias [34]; its distribution differs by medium and format [17,18], and initial fixation position, display size, visual angle, scaling, and aspect ratio all shape it. Two consequences follow. A model can transfer its bias component and fail on the image-specific component while headline metrics move little—which is why baselines and inter-observer ceilings are indispensable in any transfer evaluation [5,6]. And bias mismatch between formats is itself a transfer failure.

Temporal structure in video (evidence). Gaze in dynamic media exhibits dynamics with no static analogue [19,20]. Frame-wise application of static models misses all of it by construction; this is a change in the conditional mapping of Equation (1) that no amount of static training data repairs.

Analysis choices (evidence). Fixation-detection parameters and density-map bandwidth change the ground truth itself, and therefore change agreement statistics independently of the model; map preprocessing alone can reorder models under some metrics [6]. The consensus reporting guideline for eye-tracking studies lists the items that make such choices auditable [35]. Comparisons across studies that used different bandwidths are not comparisons of models.

4. One Model or Many? Specialization as a Pooling Problem

4.1. The Trade-Off

Sections 2 and 3 appear to argue for specialization all the way down: if categories differ, train category models. But specialization has a cost the transfer literature rarely prices in: every split of the training corpus reduces the data available to each model. With

fixation data expensive to collect, the realistic comparison is never "a category model with unlimited data versus a generalist with unlimited data"; it is "a category model trained on my category's slice versus a generalist trained on everything." Whether specialization wins then depends on whether between-category heterogeneity exceeds the benefit of pooled sample size—and on whether categories share enough structure (text pull, face pull, generic biases) that a pooled model learns each category's regularities partly from the others' data.

4.2. Three Regimes and Their Realizations

This is, formally, the pooling problem of hierarchical modeling [36]. Let θ_k denote the category-level parameter vector for category k, estimated from n_k ground-truth observations, and suppose

$$\theta_k \sim \mathcal{N}(\mu, \tau^2), \tag{2}$$

with $\tau \to 0$ recovering complete pooling and $\tau \to \infty$ recovering none. The three regimes are then points on one continuum:

- **Complete pooling—one generalist.** All categories contribute to one model ($\tau \to 0$). Maximizes effective training data and robustness to unseen material; risks averaging away category-specific structure.
- **No pooling—independent specialists.** One model per category ($\tau \to \infty$). Maximizes domain match in principle; each model sees only its slice, and rare categories get models trained on inadequate data. Operationally it multiplies everything governance must maintain—versions, monitoring, retraining, documentation—per model [29,30].
- **Partial pooling—shared structure with category-specific adaptation.** Finite τ: category-specific parameters are informed by the pooled data. This regime has several distinct technical realizations, which the literature tends to conflate: fine-tuning a pooled or pre-trained model on the category's slice [37], including parameter-efficient fine-tuning such as low-rank adaptation or adapter modules, which specialize a shared backbone under a tight per-category data budget and are now widely used for this purpose in computer vision [38]; multi-task learning with shared representations and category-specific heads [39]; mixture-of-experts models with a learned gate [40]; domain-conditioned models that take the category as an input; and hierarchical or shrinkage approaches that estimate category effects with pooling toward the grand mean [36]. They differ in data requirements and in what they can represent, but they share the property expressed by Equation (2). The strongest published fixation-prediction models are already partially pooled in this sense: they fine-tune features learned on large generic corpora with comparatively small fixation datasets [3,7,28], and the qualitative pattern at Scale 1—generic models fail on designs; design-adapted models built on generic backbones succeed [15–17]—is what partial pooling predicts.

Equation (2) turns "the answer is category-dependent" from a hedge into a claim: the right degree of pooling depends on τ—how much categories truly differ—relative to the within-category noise and the per-category data budget n_k. It yields a testable expectation: no-pooling specialists should beat the generalist only for categories that are (a) well-sampled (large n_k) and (b) atypical relative to the pooled distribution (θ_k far from μ); for typical or data-poor categories, the generalist should win—a heterogeneous, category-dependent pattern rather than a uniform victory (Figure 3). I state this as a hypothesis. No published study has tested it for gaze prediction on designed media, and the vignette in Section 5 is consistent with it but cannot confirm it.

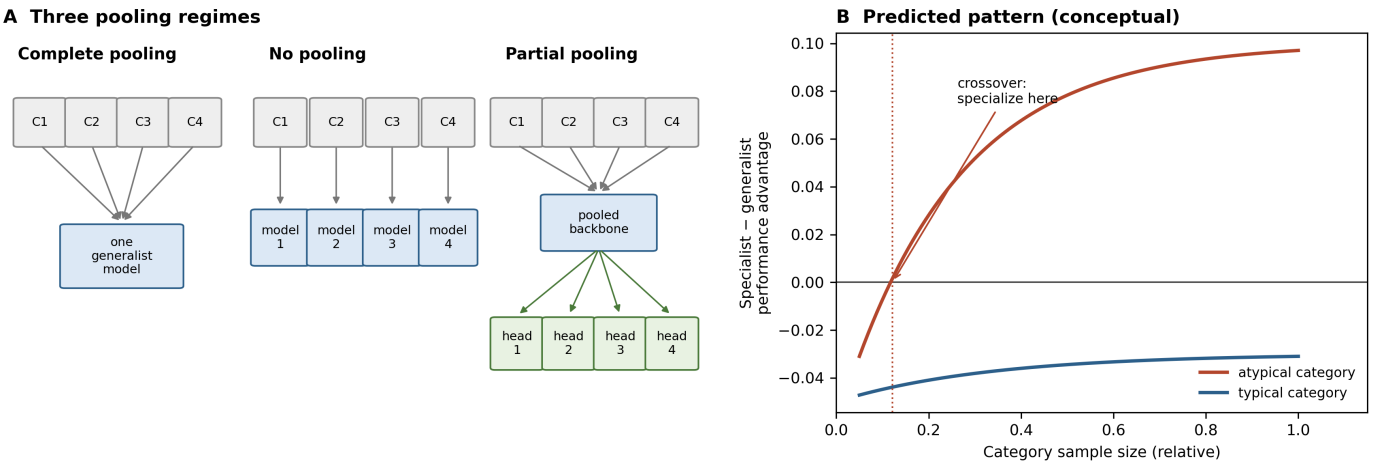

Figure 3. The pooling framework. (**A**) Three regimes as data-flow diagrams: complete pooling, no pooling, and partial pooling. (**B**) The hypothesized specialist-minus-generalist performance advantage as a function of category sample size n_k, for a category typical of the pooled distribution and for an atypical one; the crossover marks where specialization is expected to pay. The curves are conceptual, not data.

4.3. What a Test of the Hypothesis Requires

The comparison the field needs, and the one the vignette in Section 5 did not run, has a specific shape. Arms: complete pooling; specialists trained from scratch on their slice; and at least one partial-pooling realization (pooled pre-training plus fine-tuning at minimum). Controls: identical architecture, optimization budget, preprocessing, output representation, and density-map bandwidth across arms. Data: stimulus-disjoint splits that are grouped by campaign, template family, brand, source, and—for video—entire videos, because random splits of designed media leak through reformatted duplicates, campaign variants, and frames of the same clip [41,42], and pre-training corpora must be checked for overlap with test stimuli. Budget: two comparisons answer two different questions and must not be conflated—equal per-model training data (does specialization help at matched data?) and a fixed total annotation budget (how should finite data be allocated?). Estimation: repeated splits or nested cross-validation; category-level estimates with uncertainty; an explicit test of the category-by-regime interaction; and a temporally separated or source-external validation set. Metrics: as in Section 6.1, against a center-bias baseline and an inter-observer ceiling.

4.4. A Hypothesis-Driven Decision Guide

Pending that test, the analysis supports a provisional guide (Table 2), to be read as hypothesis, not rule.

Table 2. Provisional decision guide for the generalist-versus-specialist choice.

Condition	Likely Strategy	Rationale
Little target-category data; modest shift	Pooled generalist	Data volume dominates; shift small enough to absorb
Moderate target data; shared representation plausible	Partial pooling (adaptation, multi-task, or hierarchical)	Category effects estimable, informed by pooled data

Table 2. *Cont.*

Condition	Likely Strategy	Rationale
Large target data; atypical category	Specialist (or heavily weighted adaptation)	Heterogeneity exceeds pooling benefit; data sufficient
New input structure (static to video; new modality)	Dedicated architecture	Mapping change; not repairable by data allocation alone
Any of the above, deployed across markets or over time	Add the cross-cutting checks of Section 6	Population and version validity are separate questions

5. A Company Practice Vignette

This section reports unpublished internal research and development from Neurons, my company, as a vignette illustrating the operational decision process of Section 4. It is grey literature, not peer reviewed, and its evidential role is deliberately limited (Section 5.3). All statements derive from the program's final report [21], which is available for inspection.

5.1. The Program

An internal development program at Neurons (2021–2022) addressed the question of Section 4 in operational form: whether gaze-prediction models trained separately for industry verticals and content types (advertisements, websites, packaging, retail, mobile e-commerce, products; images and video) should replace a single generalist model trained on all materials [21]. Ground truth was hardware eye tracking from the company's dedicated commercial data collections. The report is a funder deliverable written for a development audience; it does not document participant numbers per stimulus, recruitment, the holdout construction used for evaluation, or whether the specialist models were trained independently or adapted from the generalist. No claims about the sample are therefore made here, and the vignette is confined to the model-development decision the report records.

5.2. What the Report Shows

The report evaluated models with in-house pixelwise statistics—regression and rank correlations between predicted and measured maps, and mean absolute error—without a center-bias baseline, an inter-observer ceiling, or the standard saliency-metric set [5,6]. Because pixelwise agreement on smoothed maps is inflated by spatial autocorrelation, smoothing bandwidth, and center bias, and because the effective sample size is not the number of pixels, the numerical values are not interpretable as accuracy evidence and are not reported here. Three qualitative observations bear on Section 4.

First, the report did not show a consistent specialist advantage. Across verticals, specialized image models varied widely relative to the generalist, some above and some below it, with no systematic relationship between the direction of the difference and the type of material; specialists for mobile e-commerce, packaging, and product imagery tended to score higher than the generalist, and specialists for advertisements and websites lower [21]. This is the heterogeneous, category-dependent pattern that Equation (2) predicts (Section 4.2), but the program's design—the generalist trained on all data, specialists on slices, with no fixed-budget arm and no partial-pooling arm—cannot separate a specialization effect from a data-volume effect.

Second, the generalist was retained on pragmatic grounds: better handling of unseen material, lower maintenance burden than a portfolio of specialists, and weaker specialist

performance in the commercially central categories [21]—an operational enactment of the governance costs of no pooling.

Third, video was handled by a dedicated model. A model trained on video eye tracking outperformed the best image model applied frame by frame, and the decisive improvement was temporal: later models learned the lag with which gaze relocates after scene cuts, achieved by varying the window of frames available for prediction, where earlier models re-fixated new targets instantaneously [21]. This should be read as architectural capability—a static model cannot represent a temporal mapping regardless of how it is trained—rather than as evidence that domain-specialized training per se is required; it is the "new input structure" row of Table 2, and a change in the conditional mapping in the terms of Equation (1). It also converges with the published finding that gaze in dynamic scenes is governed by motion and by post-cut dynamics that static features cannot capture [19,20].

The program also found that an entropy-based image-complexity index differed significantly across verticals, prompting category-specific normative ranges for derived scores even under a generalist prediction model [21]—specialization re-entering at the interpretation layer after being declined at the model layer.

5.3. What the Vignette Cannot Establish

The vignette is illustrative, not probative. It confounds specialization with data volume; it did not evaluate partial pooling as a distinct arm; its metrics lack baselines, ceilings, and standard definitions, so its magnitudes cannot be compared with the published literature; its methodological documentation is incomplete in the respects stated in Section 5.1; and I own the company that produced it. Its proper role is exclusively the one given here: an existence proof that the Section 4 trade-off, not Scale-1 doctrine, is the choice practitioners actually face, and an illustration of how that choice was made in one company. It is not performance evidence—it does not show that any regime works better than another—and it is not treated as such anywhere in this review, including in Section 7.1. The preferred upgrade—recovering the data and re-running the comparison with the design of Section 4.3—is the first item of the research agenda (Section 7.3).

6. Pre-Deployment Transfer Testing

6.1. What "Accuracy" Has to Mean

A transfer test is only as informative as the performance question it answers, and there are four distinct ones, ordered by proximity to the applied decision:

1. **Map-level agreement** between predicted and measured density. These metrics fall into distinct families that answer different questions: rank-ordering metrics that discount absolute scale and ask whether fixated locations are ranked above non-fixated ones (AUC-Judd and shuffled AUC, which also discounts center bias); distribution-similarity metrics that compare the predicted and measured maps directly as distributions (CC, SIM, KL divergence); and probabilistic calibration metrics that score a model's predicted density against a baseline or a ground-truth likelihood (information gain relative to a baseline and, for models that output probability densities, log-likelihood) [5,6]. NSS is a further location-based variant, computed at measured fixation locations rather than over the full map. Which family is appropriate depends on whether the output is a density, a binary map, or a ranking; no single metric is a universal standard.
2. **Within-asset localization:** whether the model places density on the right regions of a given asset, assessed at AOI level only when AOIs are pre-specified independently of the model output.

3. **Between-asset ranking:** whether asset-level scores derived from the maps order a set of assets as measured gaze orders them—rank correlation with uncertainty, at the within-category range users care about (Section 2.3).
4. **Decision usefulness:** whether model-assisted choices agree with, or improve on, choices made from measured gaze or from downstream outcomes.

A model can perform well on the first and fail on the third: aggregate map agreement does not guarantee that two advertisements are ranked correctly on the metric a user acts on. Commercial claims almost always report something in the first class, while applied use almost always depends on the third or fourth.

6.2. The Protocol

Table 3 states the protocol. The governing principle: transfer is established by testing on the deployment class, at the deployment scale, under deployment conditions, against baselines and ceilings, with acceptance criteria fixed in advance—never by extrapolating benchmark performance across scales. The protocol, and the decision procedure built on it (Algorithm 1), are proposed and, like the pooling hypothesis they operationalize (Section 4), untested: no study has evaluated whether following them yields better deployment decisions than the alternatives. They should be read as a recommended discipline, not a validated procedure, pending the research agenda of Section 7.3.

Table 3. Pre-deployment transfer test and reporting items for studies using predicted gaze.

Step	Requirement	Rationale/Anchor
T1	Define the intended use and deployment domain: media class, category, format, static or video, audience, device and viewing geometry, task, exposure duration	Sections 1.3 and 2; claims inherit the scale and conditions tested
T2	Assemble a held-out test set of the deployment class, disjoint from training and tuning by grouped split (campaign, template family, brand, source, whole video); disclose any known training-domain overlap. Size: determine by precision—simulate or bootstrap the metric's variance and fix the smallest practically relevant difference; as a provisional planning heuristic only, 30–50 assets	Leakage [41,42]; Section 4.3
T3	Collect measured eye tracking on the test set under conditions matching T1 (task, exposure, device), reported to the consensus minimal-guideline standard (device, sampling rate, calibration and accuracy, event classification, participants per stimulus, density-map bandwidth)	Ground-truth quality [35]
T4	Report map-level metrics appropriate to the output type (Section 6.1, class 1) and—where asset-level scores are used—between-asset rank performance (class 3), each with uncertainty intervals	Metric interpretability [5,6]
T5	Report against a center-bias baseline and an inter-observer (or gold-standard) ceiling on the same data; prefer shuffled AUC or information gain where center bias is strong	Separates generic-bias from image-specific performance (Section 3) [5,34]
T6	Report the distribution across the category's assets (not only the mean) and across audience subgroups where the intended use spans them	Range restriction [22]; cross-cutting axis (Section 2.4)
T7	Predefine acceptance: state before testing the criterion for "fit for the intended use"—for example, non-inferiority to a stated fraction of the inter-observer ceiling on the primary metric, and a minimum rank correlation for class-3 use—and report pass or fail against it	A test without a prior criterion cannot establish trustworthiness
T8	Fix and report model identity, version, and date; specify the re-test trigger on model updates; monitor after deployment	Temporal validity [29,30,43]
T9	In applied studies: report predicted metrics as predictions, cite the transfer evidence for the class and conditions used, and condition conclusions on it	Reporting discipline for prediction-model use [44]

For the specialization decision itself (Section 4.4), the protocol extends naturally, and Algorithm 1 states the procedure: run T1–T7 for the generalist on each candidate category; consider specialization (preferably a partial-pooling realization) only where the generalist fails T7 for that category; and re-run the protocol for the specialist at matched data before adopting it.

Algorithm 1 Specialization decision procedure for a deployed gaze-prediction model

Require: generalist model G; candidate categories $k = 1, \ldots, K$; for each k, a grouped held-out test set D_k with measured gaze (T2–T3); a predefined acceptance criterion A (T7); model version and date v (T8)
Ensure: a deployment decision per category and a re-test trigger
for $k = 1, \ldots, K$ **do**
  evaluate G on D_k: map-level and rank metrics with uncertainty (T4), against baseline and ceiling (T5), with distributional and subgroup checks (T6)
  if G satisfies A on D_k **then**
    deploy G for category k; record v
  else
    fit a partial-pooling adaptation M_k of G to category k at matched data (Section 4.3); evaluate M_k on a fresh grouped split
    if M_k satisfies A **then**
      deploy M_k for category k; record its version and date
    else
      escalate per Table 2: a specialist if n_k is large and k is atypical, a dedicated architecture if the input structure has changed, otherwise withhold deployment for k
    end if
  end if
end for
schedule re-test of every deployed model on any version change and at a fixed interval (T8)

6.3. Acceptance Criteria and Sample Size

Two elements of Table 3 deserve emphasis because they are the ones most often omitted. First, a predefined acceptance criterion (T7) is what converts a test into a decision: without it, any result can be read as adequate. The criterion should be expressed relative to the ceiling—for a metric m, requiring $m_{\text{model}} \geq \alpha \cdot m_{\text{ceiling}}$ for a stated α, and for class-3 use a minimum rank correlation $\rho_{\min}$ between predicted and measured asset orderings—because absolute metric values are not comparable across datasets and bandwidths [5,6]. Second, the test-set size (T2) should follow from the precision needed to detect the smallest difference that would change the deployment decision, estimated by bootstrap or simulation on pilot data; the 30–50-asset figure is a provisional heuristic for planning, not a validated minimum.

7. Discussion

7.1. What the Evidence Supports

Three conclusions are warranted now, with unequal confidence. With high confidence: Scale-1 transfer failure is established—natural-scene-trained models should not be deployed on designed media without design-domain evidence, and design-domain training reliably helps [15–18]. With moderate confidence: transfer between designed-media classes degrades, and static-to-video transfer fails for reasons of model capacity rather than data [18–20]. As a hypothesis with supporting but not confirming evidence: specialization at realistic data budgets is not automatically the remedy for heterogeneity; the generalist-versus-specialist choice is a pooling problem whose answer is category-dependent, and partial pooling is the regime most consistent with the structure of the strongest published models [3,7,28,37]. The vignette (Section 5) illustrates how this trade-off is navigated oper-

ationally, but, for the reasons given in Section 5.3, it is not independent evidence for the hypothesis and is not counted as such here. The evidence thins at exactly the scales where tools are actually used—within-category ranking and audience generalization—and the review's searches identified no published test of either for advertising material.

7.2. What the Evidence Does Not Support

Symmetry requires stating the converse. Documented Scale-1 failures do not show that current commercial tools—typically trained at least partly on designed media—fail on designed media; that is an open empirical question, answerable per tool, per category, and per condition only by tests of the Table 3 form. The vignette does not show that generalists are sufficient; its confounds preclude that reading. Nor does anything here show that the constructs attached to derived scores are measured; that is a separate validation question outside this review's scope. The honest summary is asymmetric: strong evidence that transfer cannot be assumed; weak evidence, in either direction, about how current tools actually perform at the scales and under the conditions where they are used.

7.3. Research Agenda

Five studies would change this picture. (1) The pooling experiment of Section 4.3 on designed media at both matched per-model data and fixed total budget, with a partial-pooling arm—the study the vignette almost was; recovering and re-analyzing the program's data with that design would be a legitimate first instance. (2) A within-category discrimination benchmark: can models rank professionally produced assets within a category against measured gaze, with baselines, ceilings, and predefined criteria? (3) A population and condition transfer study—markets, demographics, devices, tasks, exposures—on identical materials. (4) A video-attention benchmark for advertising with temporal metrics that score post-cut dynamics, not only frame-wise density. (5) Prospective decision-validity studies: do model-assisted choices among assets agree with, or improve on, choices made from measured gaze or downstream outcomes? All five fit this journal's scope, and none requires proprietary access.

7.4. Limitations

This is a critical narrative review; the search described in Section 1.4 is targeted and citation-chained, not systematic, so absence claims are bounded by it. The vignette is grey literature from my company, with the limitations of Section 5.3 and a direct conflict of interest; it is used to illustrate a decision process, its numerical results are deliberately not reported, and every load-bearing claim of the review stands on the published literature independently of it. No co-author without a financial interest was available to re-analyze the vignette material; the underlying report is available for inspection, which partly mitigates this. Finally, the field moves quickly; the synthesis is dated by its search date.

8. Conclusions

The validity of gaze prediction on designed media is a transfer question that recurs at progressively finer deployment scales—from natural scenes to designed media, between and within classes of designed media—and along a cross-cutting axis of observers, tasks, contexts, and time. The evidence establishes that transfer cannot be assumed, thins at exactly the scales where tools are used, and does not license the reflex conclusion that specialization is always the remedy: at finite data budgets, the generalist-versus-specialist choice is a pooling problem whose answer is category-dependent and, so far, untested. What the field needs is not a doctrine but a discipline—pre-deployment tests at the deployment scale and under deployment conditions, against baselines and ceilings, with acceptance criteria fixed

in advance; reporting that keeps predictions labeled as predictions; and the five studies of Section 7.3. Until then, trust in any gaze prediction on designed media extends exactly as far as the transfer evidence for that class, under those conditions, for that model version—and no further.

Funding: This research received no external funding. The internal program reported in Section 5 was supported by an Innobooster grant from Innovation Fund Denmark to Neurons.

Institutional Review Board Statement: This review analyzes published literature and a company technical report. It involved no new data collection from human participants, and no participant-level data were accessed or analyzed. Section 5 reports aggregate outcomes of a completed internal commercial research program. Under the Danish Committee Act (Consolidated Act No. 1268 of 28 November 2024), notification to the research ethics committee system (§ 14 of the Act) applies only to health research projects as defined in § 2, no. 1, of the Act; consumer research without a health-scientific purpose, biological material, or medical intervention falls outside the committee system's remit.

Informed Consent Statement: Not applicable. No participant-level data were collected or analyzed for this review.

Data Availability Statement: The published studies synthesized are cited in the references. The internal report underlying Section 5 [21] is available from the author upon reasonable request.

Acknowledgments: The author used Claude (Anthropic) for literature scoping, reference verification, drafting assistance, figure drafting, and language editing. All judgments and interpretations are the author's own; the author reviewed and edited all AI-assisted content and takes full responsibility for it.

Conflicts of Interest: The author is the founder and Chief Executive Officer of Neurons, a commercial vendor of gaze-prediction tools, and Section 5 reports the company's internal development program. This constitutes a direct financial interest in the subject matter. Containment: the vignette is reported with its confounds and missing documentation stated and its numerical results withheld (Sections 5.2 and 5.3); no competitor comparisons are reported; all normative claims rest on the published literature independent of the vignette; the underlying report is available for inspection; and no performance claim for any commercial model is made in the abstract or conclusions.

Abbreviations

The following abbreviations are used in this manuscript:

AOI	Area of interest
AUC	Area under the receiver operating characteristic curve
CC	Correlation coefficient (between predicted and measured density maps)
GBVS	Graph-based visual saliency
IG	Information gain
KL	Kullback–Leibler divergence
NSS	Normalized scanpath saliency
SIM	Similarity (histogram intersection of density maps)
UI	User interface
WEIRD	Western, educated, industrialized, rich, and democratic

References

1. Itti, L.; Koch, C. Computational modelling of visual attention. *Nat. Rev. Neurosci.* **2001**, *2*, 194–203. [CrossRef] [PubMed]
2. Judd, T.; Ehinger, K.; Durand, F.; Torralba, A. Learning to predict where humans look. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Kyoto, Japan, 29 September–2 October 2009; pp. 2106–2113.
3. Kümmerer, M.; Theis, L.; Bethge, M. Deep Gaze I: Boosting saliency prediction with feature maps trained on ImageNet. In Proceedings of the ICLR 2015 Workshop Track, San Diego, CA, USA, 7–9 May 2015.
4. Jiang, M.; Huang, S.; Duan, J.; Zhao, Q. SALICON: Saliency in context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1072–1080.

5. Kümmerer, M.; Wallis, T.S.A.; Bethge, M. Information-theoretic model comparison unifies saliency metrics. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 16054–16059. [CrossRef] [PubMed]
6. Bylinskii, Z.; Judd, T.; Oliva, A.; Torralba, A.; Durand, F. What do different evaluation metrics tell us about saliency models? *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 740–757. [CrossRef] [PubMed]
7. Kümmerer, M.; Bethge, M. Predicting visual fixations. *Annu. Rev. Vis. Sci.* **2023**, *9*, 269–291. [CrossRef] [PubMed]
8. Šola, H.M.; Qureshi, F.H.; Khawaja, S. Predicting behaviour patterns in online and PDF magazines with AI eye-tracking. *Behav. Sci.* **2024**, *14*, 677. [CrossRef] [PubMed]
9. Šola, H.M.; Qureshi, F.H.; Khawaja, S. AI and eye tracking reveal design elements' impact on e-magazine reader engagement. *Educ. Sci.* **2025**, *15*, 203. [CrossRef]
10. Šola, H.M.; Qureshi, F.H.; Khawaja, S. Human-centred design meets AI-driven algorithms: Comparative analysis of political campaign branding. *Informatics* **2025**, *12*, 30. [CrossRef]
11. Posner, M.I. Orienting of attention. *Q. J. Exp. Psychol.* **1980**, *32*, 3–25. [CrossRef] [PubMed]
12. Papoutsaki, A.; Sangkloy, P.; Laskey, J.; Daskalova, N.; Huang, J.; Hays, J. WebGazer: Scalable webcam eye tracking using user interactions. In Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI), New York, NY, USA, 9–15 July 2016; pp. 3839–3845.
13. Semmelmann, K.; Weigelt, S. Online webcam-based eye tracking in cognitive science: A first look. *Behav. Res. Methods* **2018**, *50*, 451–465. [CrossRef] [PubMed]
14. Borji, A. Saliency prediction in the deep learning era: Successes and limitations. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 679–700. [CrossRef] [PubMed]
15. Bylinskii, Z.; Kim, N.W.; O'Donovan, P.; Alsheikh, S.; Madan, S.; Pfister, H.; Durand, F.; Russell, B.; Hertzmann, A. Learning visual importance for graphic designs and data visualizations. In Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST), Québec City, QC, Canada, 22–25 October 2017; pp. 57–69.
16. Fosco, C.; Casser, V.; Bedi, A.K.; O'Donovan, P.; Hertzmann, A.; Bylinskii, Z. Predicting visual importance across graphic design types. In Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST), Virtual, 20–23 October 2020; pp. 249–260.
17. Leiva, L.A.; Xue, Y.; Bansal, A.; Tavakoli, H.R.; Köroğlu, T.; Du, J.; Dayama, N.R.; Oulasvirta, A. Understanding visual saliency in mobile user interfaces. In Proceedings of the 22nd International Conference on Human–Computer Interaction with Mobile Devices and Services (MobileHCI), Oldenburg, Germany, 5–8 October 2020; Article 3.
18. Jiang, Y.; Leiva, L.A.; Houssel, P.R.B.; Tavakoli, H.R.; Kylmälä, J.; Oulasvirta, A. UEyes: Understanding visual saliency across user interface types. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023; pp. 1–21. Article 285.
19. Smith, T.J.; Mital, P.K. Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. *J. Vis.* **2013**, *13*, 16. [CrossRef] [PubMed]
20. Mital, P.K.; Smith, T.J.; Hill, R.L.; Henderson, J.M. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cogn. Comput.* **2011**, *3*, 5–24. [CrossRef]
21. Neurons Inc. *Predict (prev. NeuroVision): InnoBooster Final Report*; Neurons Inc.: Taastrup, Denmark, 2022; Unpublished technical report; available from the author.
22. Hedge, C.; Powell, G.; Sumner, P. The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods* **2018**, *50*, 1166–1186. [CrossRef] [PubMed]
23. Yarbus, A.L. *Eye Movements and Vision*; Plenum Press: New York, NY, USA, 1967.
24. Rothkopf, C.A.; Ballard, D.H.; Hayhoe, M.M. Task and context determine where you look. *J. Vis.* **2007**, *7*, 16. [CrossRef] [PubMed]
25. Chua, H.F.; Boland, J.E.; Nisbett, R.E. Cultural variation in eye movements during scene perception. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 12629–12633. [CrossRef] [PubMed]
26. Henrich, J.; Heine, S.J.; Norenzayan, A. The weirdest people in the world? *Behav. Brain Sci.* **2010**, *33*, 61–83. [CrossRef] [PubMed]
27. Buolamwini, J.; Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proc. Mach. Learn. Res.* **2018**, *81*, 77–91.
28. Linardos, A.; Kümmerer, M.; Press, O.; Bethge, M. DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021.
29. Gama, J.; Žliobaitė, I.; Bifet, A.; Pechenizkiy, M.; Bouchachia, A. A survey on concept drift adaptation. *ACM Comput. Surv.* **2014**, *46*, 44. [CrossRef] [PubMed]
30. Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I.D.; Gebru, T. Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*), Atlanta, GA, USA, 29–31 January 2019; pp. 220–229.

31. Harel, J.; Koch, C.; Perona, P. Graph-based visual saliency. In *Advances in Neural Information Processing Systems 19*; MIT Press: Cambridge, MA, USA, 2007; pp. 545–552.
32. Cerf, M.; Frady, E.P.; Koch, C. Faces and text attract gaze independent of the task: Experimental data and computer model. *J. Vis.* **2009**, *9*, 10. [CrossRef] [PubMed]
33. Wang, H.C.; Pomplun, M. The attraction of visual attention to texts in real-world scenes. *J. Vis.* **2012**, *12*, 26. [CrossRef] [PubMed]
34. Tatler, B.W. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *J. Vis.* **2007**, *7*, 4. [CrossRef] [PubMed]
35. Dunn, M.J.; Alexander, R.G.; Amiebenomo, O.M.; Arblaster, G.; Atan, D.; Erichsen, J.T.; Ettinger, U.; Giardini, M.E.; Gilchrist, I.D.; Hamilton, R.; et al. Minimal reporting guideline for research involving eye tracking (2023 edition). *Behav. Res. Methods* **2024**, *56*, 4351–4357. [CrossRef] [PubMed]
36. Gelman, A.; Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*; Cambridge University Press: Cambridge, UK, 2007.
37. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems 27*; Curran Associates: Red Hook, NY, USA, 2014; pp. 3320–3328.
38. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022.
39. Caruana, R. Multitask learning. *Mach. Learn.* **1997**, *28*, 41–75. [CrossRef]
40. Jacobs, R.A.; Jordan, M.I.; Nowlan, S.J.; Hinton, G.E. Adaptive mixtures of local experts. *Neural Comput.* **1991**, *3*, 79–87. [CrossRef] [PubMed]
41. Kapoor, S.; Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **2023**, *4*, 100804. [CrossRef] [PubMed]
42. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]
43. Ramsøy, T.Z. From metric to model: A five-stage pipeline for building and sustaining valid AI predictive systems in consumer neuroscience. *Front. Hum. Neurosci.* **2026**, *20*, 1837923. [CrossRef] [PubMed]
44. Collins, G.S.; Moons, K.G.M.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; van Smeden, M.; et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **2024**, *385*, e078378. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.