DEEP-ARVIM: A PERCEPTUAL VISIBILITY METRIC AND BENCHMARK FOR AUGMENTED REALITY DISPLAYS

Anonymous authors
Paper under double-blind review

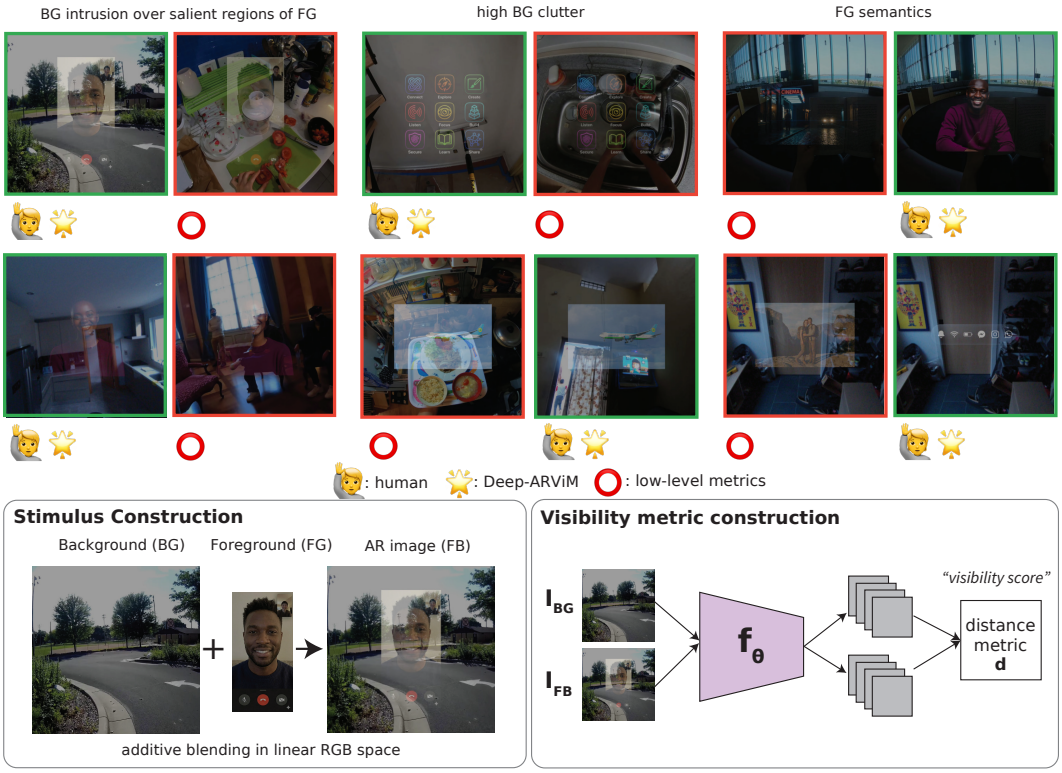

Figure 1: **In which image is the transparent foreground more visible?** Addressing this question is central for optical see-through augmented reality (OST-AR) display adaptation. To support development of a solution to this problem, we introduce Deep-ARViM, a benchmark and learned metric for AR visibility. **Top:** representative image pairs where our metric gives the correct visibility vote and low-level metrics do not. **Bottom left:** each OST-AR stimulus is constructed by additively blending a foreground with a real-world background into a *composite*. **Bottom right:** Diagram of the Deep-ARViM architecture. The metric compares the composite with the standalone background in latent space to produce a distance that we use as a visibility score.

ABSTRACT

In optical see-through augmented reality (OST-AR), virtual content is optically superimposed onto the real world, so its visibility depends jointly on the rendered foreground and the background light field. This makes visibility prediction a perceptual primitive for AR: without it, systems cannot reliably reason about display brightness, virtual object placement and rendering, or power consumption under real viewing conditions. Yet existing perceptual metrics were designed for image distortion or structural similarity. Here, we present **Deep-ARViM** (Augmented Reality Visibility Metric), to our knowledge the first learned perceptual metric explicitly designed to capture content visibility in OST-AR.

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.

We collect a large psychophysical dataset with two complementary tasks: 67,984 two-alternative forced-choice (2AFC) judgments and 21,027 visibility matching responses from more than 550 observers, spanning diverse egocentric backgrounds, foreground contents, and luminance conditions. We formulate visibility as a distance between the composited image and the standalone background, and evaluate both hand-crafted features and adapted pretrained vision backbones on their ability to predict human responses. Our best learned models achieve on par performance with an individual human observer in predicting others' 2AFC responses, recover a graded visibility scale that aligns with independent subjective ratings, and transfer zero-shot to quantitative visibility matching. Scientifically, these results suggest that AR visibility is well described by a shared latent perceptual scale that can be recovered from inexpensive binary supervision and generalized across psychophysical tasks. Practically, Deep-ARViM provides a standardized benchmark and a path toward perceptually grounded AR display adaptation.

1 INTRODUCTION

Optical see-through augmented reality (AR) glasses overlay digital content onto the user's visual environment via additive light compositing (Figure 1, bottom left), so the visibility of these virtual images is inherently dependent on their coupling with real-world backgrounds. As AR systems mature toward general-purpose computing interfaces, adapting display output to optimize visibility across diverse and dynamically changing scenes and against hardware constraints becomes a core systems challenge. Meeting this challenge requires a perceptual metric that can predict the visibility of arbitrary virtual content composited onto real backgrounds.

Classical image quality datasets (Sheikh et al., 2005; Ponomarenko et al., 2015; Lin et al., 2019) and metrics such as SSIM (Wang et al., 2004) and PSNR quantify distortions introduced by image acquisition, compression, or transmission, where a single source image is corrupted into a degraded version. Learned perceptual metrics like LPIPS (Zhang et al., 2018) and PieAPP (Prashnani et al., 2018) extend this line of work using deep features but inherit the same focus on measuring artificial deviations from the clean reference. This framing is useful for conventional image quality assessment, but AR visibility poses a different perceptual problem. We hypothesize that AR visibility is governed by global figure-ground parsing and transparent-surface perception, processes that are largely irrelevant to judging image artifacts. Therefore, accurate visibility prediction for AR requires both **a new dataset** and **a new perceptual metric** tailored to additive compositing in real scenes.

To address this limitation, we introduce Deep-ARViM, a benchmark and learned visibility metric explicitly designed for arbitrary AR content-background combinations. Our contributions are threefold:

A large-scale psychophysical benchmark for AR visibility. To ground metric learning in human perception, we collect 67K two-alternative forced-choice (2AFC) judgments and 21K visibility matching responses from over 500 observers spanning diverse foreground-background settings.

A family of learned visibility metrics. We adapt pretrained vision backbones to predict human 2AFC judgments under three training regimes: channel-weight learning, LoRA fine-tuning, and full fine-tuning. The resulting metric achieves performance on par with an actual human observer.

Evidence for a transferable visibility score. The metric trained on 2AFC judgments transfers zero-shot to the visibility matching task, showing that the metric yields a reliable visibility score that generalizes beyond a single psychophysical protocol, and can serve as a general measure of AR visibility.

Together, these contributions establish both the empirical foundation and the computational tools needed for principled AR display adaptation.

2 Related Work

Image quality and perceptual metrics: Full-reference image-quality assessment has a long history, from pixel-level error measures such as MSE and PSNR to structural similarity SSIM (Wang et al., 2003; 2004), which compares a reference image with a distorted version using local luminance, contrast, and structural statistics. LPIPS (Zhang et al., 2018) and PieAPP (Prashnani et al., 2018) start to use deep features to learn perceptual similarity (or error) from pairwise human judgments. There have been several developments on this paradigm, including DISTS (Ding et al., 2020) that brings the metric closer to the human visual system by building explicit robustness to texture resampling in deep feature space; and DreamSim (Fu et al., 2023) that extends this idea to semantic image-level similarity judgments using fine-tuned Vision Transformers (ViTs). These methods provide important architectural and conceptual precedents, but their focus on assessing distortion relative to a reference image is fundamentally different from the perceptual problem in AR, which depends on foreground-background interactions in real-world scenes.

Visibility and contrast models: Classical visibility models (Daly, 1992; Lubin, 1995) predict detectable image differences using contrast sensitivity (Campbell & Robson, 1968) and visual masking. Psychophysical studies further characterize how luminance, contrast, and background structure affect target detection in natural scenes (Sebastian et al., 2017). Fukiage & Oishi (2023) and Kim et al. (2025) take a similar approach to measuring supra-threshold image differences and validate their model on complex images. We likewise focus on images that are already visible to varying degrees, and in addition to re-evaluating the effects of these hand-crafted cues on more realistic AR scenes, we extend to the semantic interactions between foreground and background using deep networks.

AR display evaluation: Prior work on AR display evaluation has examined specific perceptual phenomena, including accommodation-vergence conflict (Kramida, 2016), text legibility (Gabbard et al., 2006; 2018), the brightness required for virtual objects to appear opaque (Liu et al., 2022), appearance matching between virtual and real objects (Kruijff et al., 2010; Murdoch, 2020; Zhang, 2022), and display artifacts of AR videos (Chapiro et al., 2024). Our work develops a general-purpose visibility metric from inexpensive pairwise visibility judgments and demonstrates that the same score can transfer to subjective ratings and quantitative tasks.

3 Perceptual Dataset Collection

This section describes perceptual judgment dataset collection, a key prerequisite for metric learning that provides the ground truth for subsequent training and evaluation procedures. The dataset comprises two complementary tasks: a two-alternative forced choice (2AFC) task yielding ordinal visibility preferences and a visibility matching task yielding quantitative visibility measurements.

We performed the main data collection on Prolific (Prolific, 2026) (https://www.prolific.com) using simulated AR stimuli presented on conventional displays (see Appendices A.1 and A.2 for details). This design allows us to collect the scale of perceptual judgments needed for fine-tuning deep metrics, which would be prohibitively costly to obtain through in-person experiments using specialized optical setups. We conducted a separate haploscope study under real additive light to verify our choice, and the results showed: 1) strong agreement with web judgments and, 2) our web-trained metric can predict haploscope responses without any further training (Appendix D).

3.1 Stimulus design

Each simulated AR image in our dataset is constructed from a background scene, a foreground element, and a relative luminance level between the two (see Figure 1, bottom left).

Backgrounds: To ensure that our backgrounds reflect the statistics of egocentric viewpoints, as would be the case for OST-AR headsets, we sourced backgrounds from two large-scale egocentric video datasets: Aria Everyday Activities (Lv et al., 2024) and Ego4D (Grauman et al., 2022). These span hundreds of hours of daily activities across a wide range of tasks where people may be interacting with their AR display.

Foreground content: We chose foreground images to span common AR elements, including natural images of objects, portraits, and landscapes, as well as application interfaces such as messaging threads, interactive icons and maps. Examples can be found in Appendix A.3. All foreground images were generated using Gemini 3 Pro (Google, 2025).

Compositing: To simulate optical see-through AR display physics, we composite foreground onto background via additive blending in linear RGB space (see Figure 1, bottom left for an example), where the summation of light sources accurately reflects the underlying physics. The background luminance is fixed at a relative level of 1.0, and the foreground is scaled by a luminance multiplier $\ell \in \{0.0, 0.1, \ldots, 2.0\}$. The composited image is then:

$$I_{\text{FB}} = \text{sRGB}\left(\frac{I_{\text{BG}}^{\text{linear}} + \ell \cdot I_{\text{FG}}^{\text{linear}}}{\ell_{\max}}\right), \tag{1}$$

where $\ell_{\max} = 3.0$ is a normalization constant ensuring all images are normalized by the same maximum luminance.

3.2 The 2AFC task

The top panel in Figure 1 shows six examples of real trials, where a participant views two composited AR images side by side and selects the one in which the **foreground** is more visible. To enhance these ordinal-scale data, participants were then asked to rate the visibility of their chosen image on a five-point scale: very well, well, fairly, poorly, or very poorly (see rating rubric in Appendix A.5).

Each participant completes one "experiment set" consisting of approximately 120 trials. The trial structure was designed to to disentangle the effect on visibility from the foreground-background interactions (see Table 1). Each set included 6 "catch trials" where the two images have the same foreground and background image content but one has a much higher foreground luminance and should always be chosen. Approximately 10% of participants failed more than one catch trial and are excluded from analysis. Ultimately, each unique image pair has responses from at least 5 participants (see Table 2 for summary statistics).

Table 1: Breakdown of the 120 trials completed by a single participant.

Trial type	**Portion**	**Purpose**
Same FG, different BG	30%	Isolate the effect of BG on visibility.
Same BG, different FG	30%	Isolate the effect of FG on visibility.
Different FG and BG	30%	Test generalization across conditions.
Catch trials	5%	Attention checks.
Repeated trials	5%	Assess within-subject consistency.

For each trial, the foreground luminance is randomly drawn from $\{0.5, 1.0\}$. These two levels were selected based on a pilot study, which showed that the two values lie in the transition regime where changes in luminance most strongly affect perceived visibility. At $\ell = 0.5$, foregrounds are visible with careful inspection, but the background interferes substantially. At $\ell = 1.0$, visibility is mostly fine.

3.3 The visibility matching task

To obtain a quantitative and fine-grained measure of AR visibility beyond binary preference judgments, we designed the visibility matching task shown in Figure 2. In each trial, the reference image has a fixed foreground luminance ($\ell = 1.0$), and the test image has foreground luminance that the participant can adjust over 21 discrete levels in $\{0.0, 0.1, \ldots, 2.0\}$. Participants were asked to move the slider until the foregrounds in the two images appear equally visible. If the maximum luminance is still insufficient to match the reference, the participant clicks "not visible enough." To allow direct comparability with the 2AFC judgments, the image pairs used in the visibility matching task were randomly subsampled from the 2AFC task, with 40 trials in each experimental set. Additional quality-control procedures for the visibility matching task are described in Appendix A.6, and summary counts after filtering are reported in Table 2.

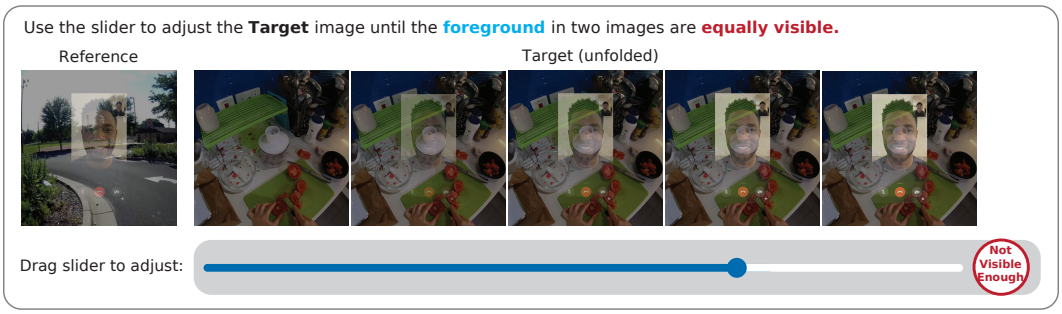

Figure 2: An example trial of the visibility matching task.

Table 2: Summary statistics of the Deep-ARViM perceptual dataset.

Statistic	**2AFC task**	**visibility matching task**
Participants	550 (10.1% excluded)	526 (14.1% excluded)
Response counts	67,984 judgments on 11,156 pairs	21,027 responses on 3,672 pairs
Response format	Left / Right	Luminance value / not visible enough

4 Learning a Visibility Score for AR

4.1 Visibility formulation

In AR, visibility is not a property of the foreground (I_{FG}) or background (I_{BG}) alone: It emerges from their interaction in the composite image (I_{FB}) formed on the retina. Let $d(\cdot, \cdot)$ be a distance function that describes how different two images are to one another. A natural formulation of visibility is to measure how much the perceptual response is altered before and after the foreground is added. Then, $d(I_{BG}, I_{FB}) \approx 0$ implies the composite is indistinguishable from the original background, and thus the added foreground is perceptually "invisible", whereas $d(I_{BG}, I_{FB}) >> 0$ indicates greater foreground visibility.

Another design choice is to find a feature extractor f_θ that captures information most relevant to this perceptual problem while balancing predictive accuracy and computational cost, yielding $d(I_{BG}, I_{FB}; f_\theta)$.

Low-level features. We evaluate f_θ measuring local luminance, CIE Lab color, Fourier spectrum, PSNR, SSIM, foreground pixel count, background clutter (Rosenholtz et al., 2007) (see Appendix B.1 for definitions). To provide a stronger baseline, we linearly combine these low-level metrics with learnable weights optimized for predicting human judgments.

Deep features. To capture semantic foreground-background interactions beyond the hand-crafted cues, we also instantiate f_θ with deep representations. We evaluate seven pretrained backbones spanning several architectures and pretraining paradigms: supervised convolutional networks VGG-19 (Simonyan & Zisserman, 2015) and ResNet-50/101 (He et al., 2016), self-supervised vision transformers DINOv2 ViT-S/14 and L/14 (Oquab et al., 2023), a multimodal transformer CLIP ViT-L/14 (Radford et al., 2021), and an object detection model DETR (Carion et al., 2020). This set allows us to test how AR visibility depends on architecture, model scale, and supervision.

Concretely, given a DNN backbone f_θ, we extract representations from a small set of representative layers $f_\theta^{(\ell)}$ spanning early, intermediate, and high-level processing. We normalize the layer activations by the channel dimension, denoted by $y_\theta^{(\ell)} \in \mathbb{R}^{H^\ell \times W^\ell \times C^\ell}$, then scale each channel by non-negative weights $w^{(\ell)} \in \mathbb{R}^{C_\ell}_{\geq 0}$. We choose d to be weighted L_2 distance on normalized activations and average spatially and over channels across all layers:

$$d(I_{BG}, I_{FB}; f_\theta) = \frac{1}{|\Omega|} \sum_{\ell=1}^{L} \sum_{c=1}^{C_\ell} \sum_{h,w} \|w^{(\ell)} \odot \big(y_\theta^{(\ell)}(I_{FB}) - y_\theta^{(\ell)}(I_{BG})\big)\|^2, \tag{2}$$

where Ω indexes spatial locations and all channels. For CLS-style outputs, the same expression applies without spatial averaging.

4.2 Learning the visibility metric

To better align the perceptual metric with human judgments, we train $\{w^\ell, \theta\}$ directly on the 2AFC data. For each trial, we take the human majority vote as the training signal. Let $y^* \in \{0, 1\}$ denote the majority label, where $y^* = 1$ means that the human majority judged the first image to have the more visible foreground. The model also gives its prediction as $\hat{y} = \mathbf{1}[s_1 > s_2]$, where s_1 and s_2 are the visibility scores of the two composite images. To promote alignment, we use the binary cross-entropy loss:

$$\mathcal{L}_{\{w^\ell,\theta\}} = -\, y^* \log \sigma(s_1 - s_2) - (1 - y^*) \log\big(1 - \sigma(s_1 - s_2)\big). \tag{3}$$

To address the challenge of tuning the large number of model parameters, we study three adaptation regimes that encompass lightweight calibration, parameter-efficient adaptation, and end-to-end tuning. The first keeps the backbone frozen and learns only the nonnegative channel weights $\{w^{(\ell)}\}$, giving an LPIPS-style calibration baseline (Zhang et al., 2018). The second combines channel-weight learning with LoRA[1] (Hu et al., 2021), which adapts the backbone by training only a small fraction of its parameters. The third performs full parameter fine-tuning. Representative layers and training hyperparameters are summarized in Appendix C.

5 Predicting the 2AFC Decisions

5.1 Evaluation metric and human reference levels

Human judgments in the 2AFC task vary as observers do not always agree on which image has the more visible foreground. We therefore evaluate models against the majority vote while accounting for inter-observer consistency (an empirical distribution is shown in Appendix A.4). For stimulus pair i, let $p_i \in [0.5, 1]$ denote the fraction of observers who chose the majority label; a predictor receives score p_i if it agrees with the majority and $1 - p_i$ otherwise. This metric admits two natural human baselines that bound expected model performance. A predictor that always matches the majority provides the upper bound:

$$\text{Accuracy}_{\text{majority}} = \frac{1}{N} \sum_{i=1}^{N} p_i = \bar{p}, \tag{4}$$

whereas a randomly sampled human observer that agrees with the majority with probability p_i and disagrees with probability $1 - p_i$ would give

$$\text{Accuracy}_{\text{individual}} = \frac{1}{N} \sum_{i=1}^{N} \left[p_i^2 + (1 - p_i)^2\right]. \tag{5}$$

5.2 Model results

Figure 3 (left) shows that:

1. Low-level features explain a substantial portion of visibility judgments. Their predictive performance makes these simple metrics attractive candidates when computational time and cost are heavily constrained. Among individual features, the relative luminance between the composite and the background is the strongest predictor, and a learned

[1] We apply LoRA only to transformer backbones because it adapts attention modules.

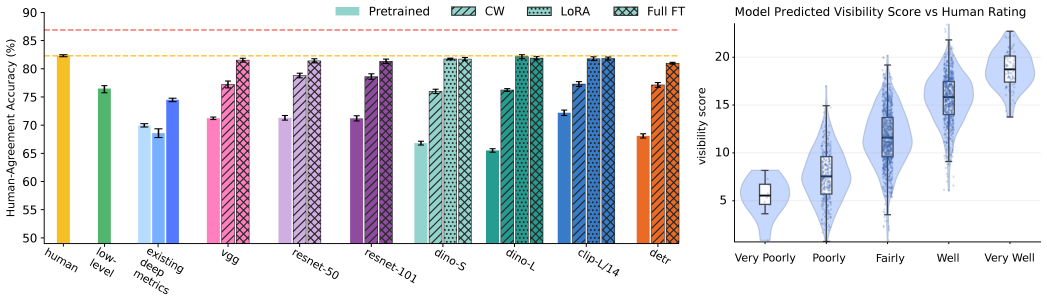

Figure 3: **Performance on the 2AFC task and recovery of a graded visibility scale. Left:** inter-observer-consistency-weighted accuracy on the benchmark, with red dashed line representing the ground truth. Gold denotes the individual-subject reference level, and green denotes the learned low-level baseline; existing deep metrics include LPIPS, DISTS, and DreamSim, respectively. For each backbone, the remaining bars show the pretrained representation and the available adaptation regimes (CW, LoRA, and full fine-tuning). Error bars are computed from different train/val splits. **Right:** distribution of visibility scores for images grouped by the five-level subjective visibility rating (from the best model). Higher predicted visibility scores correspond systematically to higher ratings.

combination of low-level cues improves this only slightly (see Appendix B.2 for details). However, the low-level baselines still fall noticeably short of the human reference levels (red and yellow dashed lines); examples of representative mistakes are indicated by the red circles under the images in Figure 1, showing that accurate visibility prediction also requires high-level (foreground saliency and semantics) and mid-level (background structure) understanding.

2. Off-the-shelf perceptual metrics trained on image distortions (Zhang et al., 2018; Ding et al., 2020) and mid-level similarity (Fu et al., 2023), as well as deep representations pretrained on general vision tasks perform poorly on our visibility benchmark, notably worse than the low-level metrics. This is strong evidence that learned features for generic image encoding do not automatically transfer to the AR visibility problem.

3. After training on human 2AFC judgments, however, performance improves substantially and consistently across backbone families and adaptation regimes. All models are trained on the training split and evaluated on a disjoint validation split. The best model (DINOv2 ViT-L/14 with LoRA) matched the expected accuracy of an individual human observer (82.6% model vs. 82.3% human). Performance is broadly similar across model architectures, model sizes, and pretraining tasks, suggesting that the main limitation is not the expressive power of deep features but the lack of alignment between those features and human AR perception.

The learned metric subsumes low-level cues. To help understand what properties of the DNN representations were capturing image visibility, we tested whether combining their prediction with the family of low-level cues can improve performance over the DNN alone. We found low-level cues add little information to the network metric, suggesting that the network already subsumes all these features for visibility prediction. In contrast, low-level metrics can be significantly improved when combined with network layers, and the gain grows monotonically with layer depth, revealing a need beyond generic image statistics and for semantic understanding. Detailed results, together with network ablations, can be found in Appendix E.

Where does the model make mistakes? Figure 4 presents counts of model correct (gray) and incorrect (red) responses (defined by agreement with the human majority vote) as a function of decision margin, defined by $|s_1 - s_2|$, and separated by different levels of inter-observer consistency. Larger margins indicate a stronger preference for one candidate, whereas smaller margins indicate near-indifference.

Consistent with what we would want from a good metric, errors are concentrated at small margins across all consistency bands. This is an important qualitative result: when the model fails, it is usually near the decision boundary rather than through confident mistakes. A recurring failure mode of our model occurs on weakly structured backgrounds, such as nearly uniform regions. These backgrounds induce weak baseline activations in DNNs, so foreground objects can create a disproportionately large shift in layer activations and lead to visibility overestimation.

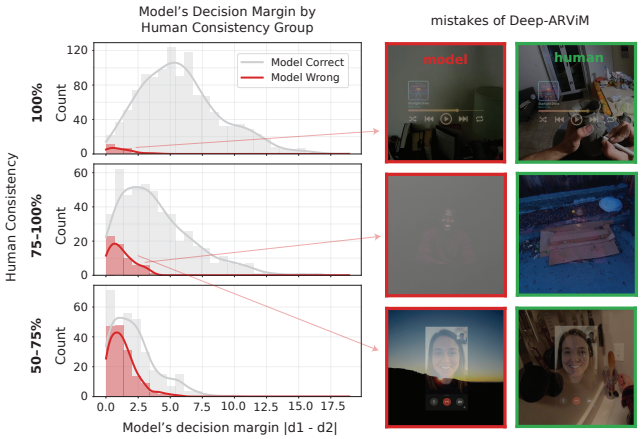

Figure 4: Decision-margin analysis and representative failure cases of the best model. Gray and red histograms show correct and incorrect predictions, respectively, grouped by 2AFC inter-observer consistency.

Does the learned score capture visibility ratings? Even though the subjective quality rating is never used in training, Figure 3 (right) shows a strong correlation between model predicted visibility scores and the discrete ratings. This result suggests the strong generalizability of our metric from binary supervision to perceptual scales of visibility. In practical terms, the score is easily interpretable and supports simple rules of thumb; for example, images with scores above about 15 are reliably rated "Well" or better.

6 Can Binary 2AFC Visibility Judgments Generalize to Quantitative Prediction?

We next ask whether our family of metrics, trained only on pairwise comparisons, can reliably predict the visibility of content in a single AR image. We test this in a zero-shot setting using data from the visibility matching experiment. The metric never sees this task during training. Success in this setting would therefore show that the learned score captures aspects of visibility that human observers re-use across related tasks rather than merely a task-specific heuristic or strategy linked to the 2AFC structure of the training data.

For each trial in the visibility matching experiment, the reference image defines a target visibility score s^*. We evaluate the same metric on the test image at every discrete luminance level $\ell \in \{0.0, 0.1, \ldots, 2.0\}$ and find the luminance whose score is closest to reference, i.e., $\hat{\ell} = \arg\min_\ell |d(I_{BG}, I_{FB}(\ell); f_\theta) - s^*|$. If $\ell = 2.0$ still cannot reach the reference visibility, the model predicts "not visible enough." We compare these zero-shot predictions against the human median responses to the same image pairs.

Figure 5 (left) shows that the learned models achieve substantially lower mean absolute error than both off-the-shelf pretrained features and the low-level baseline, showing that the learned metric generalizes beyond the task on which it was trained. Comparing the correlation between model-predicted luminance and human median responses reveals the same pattern (Appendix F). Figure 5 (right) further shows that models with better 2AFC accuracy tend to attain lower matching error, indicating that the two tasks are driven by a common underlying signal. Together, these results suggest that binary supervision is sufficient to recover a shared latent visibility scale that transfers to quantitative prediction on individual AR images.

7 Discussion

AR differs fundamentally from traditional opaque displays: the retinal image is a composite of rendered content and a background determined by the environment and user behavior. Consequently,

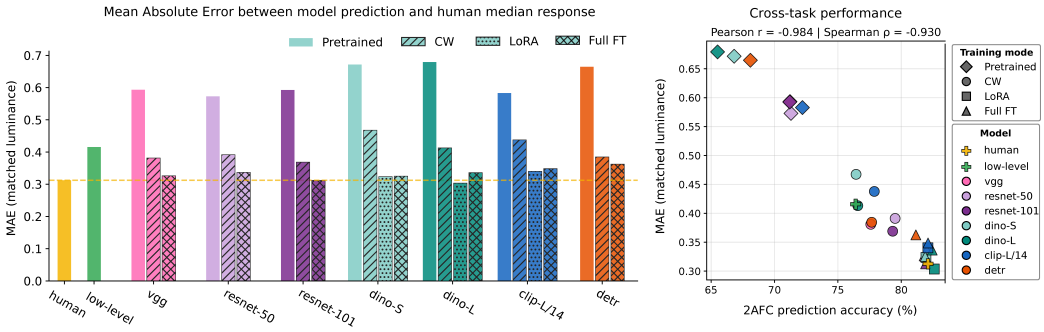

Figure 5: Zero-shot transfer from binary 2AFC visibility judgments to the visibility matching task. **Left:** mean absolute error between the model predicted luminance and the human group-median matched luminance, evaluated on the validation split. **Right:** relationship between 2AFC accuracy and mean absolute error across backbone families and training modes, showing strong correlation between the two tasks' performance.

adapting an AR display in real time (i.e., its brightness, tone mapping, content positioning etc.) is required to ensure that content remains visible across the joint distribution of real-world tasks, environments, and content types. We address the lack of visibility measurements for this setting with a dataset and benchmark of human-judged visibility on realistic egocentric foreground-background pairs. Further, we show that Deep-ARViM, our metric trained on this dataset, (i) predicts pairwise visibility choices as well as human observers, (ii) correlates with independent subjective ratings, (iii) transfers zero-shot to a visibility matching task, and (iv) outperforms low-level image metrics such as those modeling foreground-background contrast perception. This consistency suggests that AR visibility is a coherent perceptual quantity amenable to scalar estimation.

Figure 11 in Appendix G illustrates one potential application: using the learned metric to evaluate candidate placements within a scene and identify locations where the same AR content would be most visible. We also argue that our metric can fit under the compute budget of commercial smart glasses (see Appendix H).

Despite these successes, it is important to acknowledge several limitations of our approach. **Limited luminance range:** Data were collected on displays spanning 300–500 cd/m², far narrower than the human visual system's adaptation range (10^{-6}–10^{8} cd/m²) (Kunkel et al., 2016). Incorporating models of light adaptation into Deep-ARViM will be necessary for deployment across real-world conditions. **Simulated compositing**. The dataset uses simulated rather than optically-accurate compositing and does not account for spectral characteristics or optical aberrations of real AR systems; online collection also limits calibration control. Although we validated the metric under real additive light in a haploscope (Appendix D), future work should evaluate performance on physical displays under dynamic conditions. **Depth and motion:** Stimuli in our experiment were static and coplanar, while in real AR headsets, relative depth is expected to affect visibility via binocular fusion limits (e.g., Panum's fusional area) and ocular defocus, and motion cues are expected to increase salience. Our metric likely represents a lower bound on visibility with respect to these cues. **Task-specificity:** Different tasks will require different visibility trade-offs (e.g., overly bright AR instructions for a physical task can obscure the real-world scene and impede task execution), and there are broader safety concerns when AR content occludes critical environmental information. Our metric provides a reliable visibility score for AR content that can inform such trade-offs, but determining the appropriate task-specific balance remains an open problem.

Applied to display control, our metric can enable principled optimization and rendering policies under constraints such as visibility, power, and comfort, given only spatially-aligned images of content and background. Compared to low-level image metrics, the successes and errors of our metric appear to better align with humans, capturing effects like multi-scale spatial masking and semantic-level goals. We hope future work can build on our contributions both by improving our ability to predict content visibility in AR and by leveraging Deep-ARViM to explore this rich tradespace.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for the following: generate synthetic data sets, implement methods, clean and reformat dataset, support qualitative and thematic data analysis.

We have not used generative AI tools for the following: help develop theoretical models or conceptual frameworks, design or provide feedback on research methodology or experiments, interpret results.

These are not applicable to this work: formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, assist with translation.

Additionally, we used generative AI tools for the following: create or modify scientific figures or images, suggest experimental parameters, create or edit software code, draft parts of a research paper, summarize or analyse existing literature, brainstorming, sourcing/searching for information, edit a research paper to improve readability, identify relevant literature, format references, suggest a structure for a research paper, propose a title or keywords for a research paper.

We have reviewed all AI-assisted work. We carefully inspected the images that LLM generated; LLM-generated code was verified and tested for correctness by at least 1 author; we proof-read and manually edited the entire paper; we carefully read and verified all the reference papers LLM suggested. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Fergus W Campbell and John G Robson. Application of Fourier analysis to the visibility of gratings. *The Journal of Physiology*, 197(3):551–566, 1968.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pp. 213–229, 2020.

Alexandre Chapiro, Dongyeon Kim, Yuta Asano, and Rafał K. Mantiuk. AR-DAVID: Augmented reality display artifact video dataset. *ACM Transactions on Graphics*, 43(6), 2024.

Scott J Daly. Visible differences predictor: an algorithm for the assessment of image fidelity. *Proc. SPIE*, 1666:2–15, 1992.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *CoRR*, abs/2004.07728, 2020. URL https://arxiv.org/abs/2004.07728.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *NeurIPS*, 2023.

Taiki Fukiage and Takeshi Oishi. A content-adaptive visibility predictor for perceptually optimized image blending. *ACM Transactions on Applied Perception*, 20(1):1–29, 2023. doi: 10.1145/3565972. Article 3.

Joseph L Gabbard, J Edward Swan, and Deborah Hix. The effects of text drawing styles, background textures, and natural lighting on text legibility in outdoor augmented reality. *Presence*, 15(1): 16–32, 2006.

Joseph L Gabbard, Divya Gupta Mehra, and J Edward Swan. Effects of ar display context switching and focal distance switching on human performance. *IEEE transactions on visualization and computer graphics*, 25(6):2228–2241, 2018.

Google. Gemini 3 pro. https://gemini.google.com, 2025. Accessed: Nov, 2025.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Dongyeon Kim, Maliha Ashraf, Alexandre Chapiro, and Rafał K. Mantiuk. Supra-threshold contrast perception in augmented reality. *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025. doi: 10.1145/3757377.3763824.

Gregory Kramida. Resolving the vergence-accommodation conflict in head-mounted displays. *IEEE Transactions on Visualization and Computer Graphics*, 22(7):1912–1931, 2016.

Ernst Kruijff, J Edward Swan, and Steven Feiner. Perceptual issues in augmented reality revisited. In *ISMAR*, pp. 3–12, 2010.

T. Kunkel, S. Daly, S. Miller, and J. Froehlich. Chapter 15 - perceptual design for high dynamic range systems. In Frédéric Dufaux, Patrick Le Callet, Rafał K. Mantiuk, and Marta Mrak (eds.), *High Dynamic Range Video*, pp. 391–430. Academic Press, 2016. ISBN 978-0-08-100412-8. doi: https://doi.org/10.1016/B978-0-08-100412-8.00015-2. URL https://www.sciencedirect.com/science/article/pii/B9780081004128000152.

Hanhe Lin, Vlad Hosu, and Dietmar Saupe. KADID-10k: A large-scale artificially distorted IQA database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–3, 2019.

Jingyu Liu, Akshay Jindal, Claire Mantel, Soren Forchhammer, and Rafal K. Mantiuk. How bright should a virtual object be to appear opaque in optical see-through AR? In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 676–684, 2022.

Jeffrey Lubin. A visual discrimination model for imaging system design and evaluation. *Vision Models for Target Detection and Recognition*, pp. 245–283, 1995.

Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, et al. Aria everyday activities dataset. *arXiv preprint arXiv:2402.13349*, 2024.

Michael J. Murdoch. Brightness matching in optical see-through augmented reality. *Journal of the Optical Society of America A*, 37(12):1927–1936, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30:57–77, 2015.

Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. PieAPP: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1808–1817, 2018.

Prolific. Prolific [online participant recruitment platform], 2026. URL `https://www.prolific.com`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Ruth Rosenholtz, Yuanzhen Li, and Lisa Nakano. Measuring visual clutter. *Journal of vision*, 7(2): 17–17, 2007.

Stephen Sebastian, Jared Abrams, and Wilson S Geisler. Constrained sampling experiments reveal principles of detection in natural scenes. *Proceedings of the National Academy of Sciences*, 114 (28):E5731–E5740, 2017.

Hamid R. Sheikh, Zhou Wang, Lawrence K. Cormack, and Alan C. Bovik. Live image quality assessment database release 2. `https://live.ece.utexas.edu/research/quality`, 2005. Accessed 2026.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The thrity-seventh asilomar conference on signals, systems & computers, 2003*, volume 2, pp. 1398–1402. Ieee, 2003.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Lili Zhang. *Lightness, Brightness, and Transparency in Optical See-Through Augmented Reality*. PhD thesis, Rochester Institute of Technology, 2022.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, 2018.

A Details On Dataset Collection

A.1 Participant recruitment and ethics

We recruited all participants via Prolific with minimal exclusion criteria imposed: normal or corrected-to-normal vision; aged 18–69 years. All participants were informed about the tasks they would perform in the experiment, and explicit consent was obtained prior to participation. Participants were paid at a rate of $16 per hour, which is above Prolific's recommended rate. We collected no personally identifiable information, and the study was approved by the organization's research ethics review process.

A.2 Hardware

Participants were asked to use their own desktop computer to complete the task, and they were asked to adjust display brightness to the maximum throughout the experiment. Due to the nature of the data collection procedure, there will be variability in display characteristics and viewing conditions.

A.3 AR foreground examples

Figure 6 shows representative foreground images used in our study. To cover a broad range of content that may appear in everyday AR use, the set spans both natural images and common AR interface elements, including objects, portraits, landscapes, messaging, maps, and application-style graphics. To contol for any effect of relative image size, each image was rescaled to occupy 25% of the background area and placed at the center of the background.

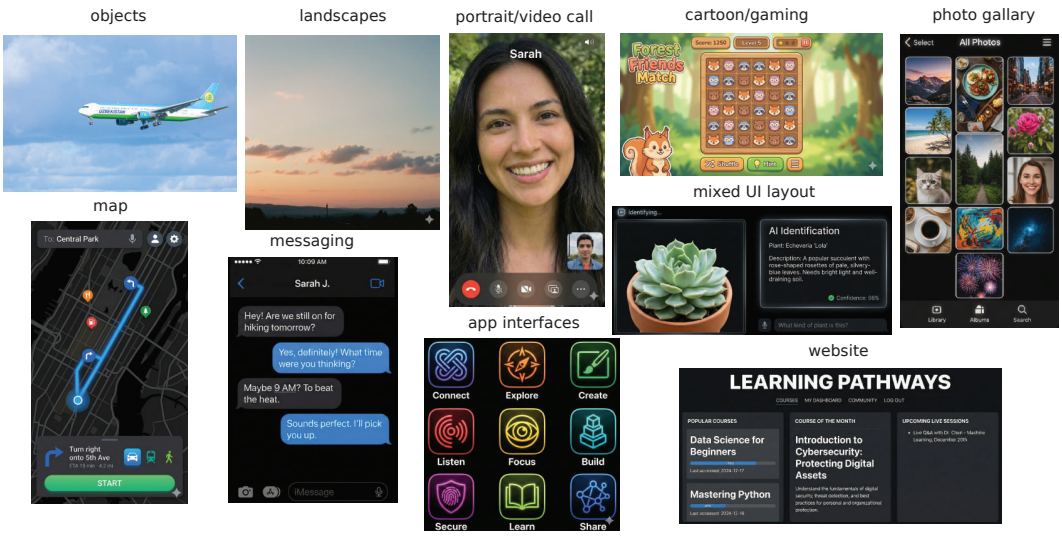

Figure 6: Representative foreground images used in the Deep-ARViM dataset. The examples span natural visual content and common AR interface elements, illustrating the diversity of foreground types used in our perceptual experiments.

A.4 2AFC consistency distribution

Figure 7 reports the empirical distribution of inter-observer consistency for image pairs from the 2AFC task, where each image pair received at least 5 responses. Most image pairs received unanimous votes.

A.5 Subjective visibility rating rubric

After selecting the more visible image in the 2AFC task, participants rated the visibility of the chosen image using the following five-level rubric:

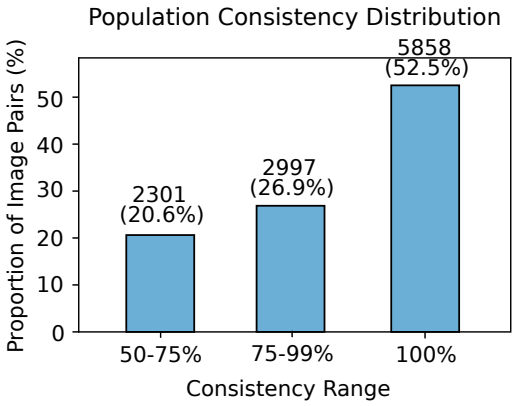

Figure 7: Inter-observer consistency in the 2AFC task.

Table 3: Distribution of subjective visibility ratings in the 2AFC task after subject filtering, ordered from very well to very poorly. Percentages are computed over all 67,984 valid ratings.

Rating	Count	Percentage
Very well	7,933	11.67%
Well	19,667	28.93%
Fairly	24,094	35.44%
Poorly	13,271	19.52%
Very poorly	3,019	4.44%
Total	67,984	100.00%

- **Very well:** the observer can see the foreground clearly with essentially no interference from the background.
- **Well:** the observer can identify the foreground easily, with only faint interference from the background.
- **Fairly:** the observer can identify the foreground with moderate effort, and the background is no longer strongly distracting.
- **Poorly:** the observer can mostly make out the foreground with careful inspection, but the background interferes substantially.
- **Very poorly:** the observer can tell that a foreground is present, but it is almost impossible to identify.

The defintion of each rating was available on screen throughout the experiment. Table 3 reports the empirical distribution of these subjective visibility ratings. We obtained a wide coverage of ratings that are slightly skewed towards the positive side.

A.6 Quality control for the visibility matching task

Because the visibility matching task requires a continuous adjustment rather than the more sharply defined binary comparison used in the 2AFC task, its responses are inherently noisier and require additional filtering beyond the 2AFC subject filtering. We apply two additional stages. First, we remove zero-luminance responses, since a choice of $\ell = 0$ implies that no foreground is present at all in the test image and thus should never be the expected visibility match. We additionally exclude subjects who produce more than three such responses. Second, we exclude subjects whose leave-one-out Pearson correlation with the group median falls below $r = 0.2$, filtering out participants whose responses are poorly aligned with the overall judgments of the retained cohort.

B LOW-LEVEL BASELINES

B.1 LOW-LEVEL METRIC DEFINITIONS

The low-level visibility metrics are very important baselines for our benchmark because they isolate how far simple hand-crafted cues can go before learned representations become necessary.

Luminance ratio family. Our luminance baseline is given by the mean luminance of the composite over that of the background: $d(I_{BG}, I_{FB}) = \text{mean}_{(h,w)\in\text{fg}} \frac{\text{mean}(L_{FB})}{\text{mean}(L_{BG})}$, where larger ratios indicate that the composite stands out more strongly from the background. Images are converted from sRGB to linear RGB, then to luminance using $L = 0.2126R + 0.7152G + 0.0722B$. We also evaluated other luminance variants, including max- or median-based statistics and versions that use a broader background region, but these performed worse and are not used in the final combined baseline.

CIE Lab color distance. This metric quantifies chromaticity difference between the composite and the background in the perceptually uniform CIE Lab color space. Both images are converted to Lab, and the Euclidean distance is computed in the a^*b^* chrominance plane (ignoring L^*), averaged over foreground pixels: $d(I_{BG}, I_{FB}) = \text{mean}_{(h,w)\in\text{fg}} \sqrt{(a^*_{BG} - a^*_{FB})^2 + (b^*_{BG} - b^*_{FB})^2}$, where a higher value indicates better visibility.

Fourier spectral distance. This metric compares the spatial-frequency structure of the composite and the background images. After conversion to luminance, both images are smoothed by a 2D Hann window and then transformed with a 2D FFT. We computed the RMS difference between magnitude spectra after removing the DC component. Direction: larger distance $\rightarrow$ more visible.

PSNR and SSIM. We also evaluate two standard full-reference image-similarity measures: PSNR measures pixelwise fidelity through mean-squared error, while SSIM measures structural similarity using local luminance, contrast, and correlation statistics. For both metrics, larger values indicate that the composite remains more similar to the background and thus corresponds to lower visibility.

Foreground pixel count. The simplest baseline counts nonzero pixels in I_{FG}—pixels where the mean RGB value exceeds 0.03. This metric captures the intuition that larger overlays are more visible: $\text{count}(I_{FG}) = |\{(h, w) : \frac{1}{3}(R_{h,w} + G_{h,w} + B_{h,w}) > 0.03\}|$. Direction: larger count $\rightarrow$ more visible.

Background clutter (Feature Congestion). This metric implements the Feature Congestion measure of Rosenholtz et al. (Rosenholtz et al., 2007), which quantifies the visual clutter level of the background I_{BG} based on color, orientation, and luminance contrast. Direction: larger clutter $\rightarrow$ less visible.

B.2 LOW-LEVEL METRIC PERFORMANCE

Figure 8 reports the inter-observer-consistency-weighted accuracy of the individual low-level metrics together with the learned linear combination. Among the individual baselines, a simple luminance-based measure—the ratio between mean composite and mean background luminance—is the strongest predictor, and a learned linear combination of low-level cues improves slightly further. This pattern shows that visibility in OST-AR is strongly constrained by basic image properties, with luminance contrast playing the dominant role. At the same time, even the best low-level baselines remain clearly below the human reference levels, indicating that visibility is not reducible to a short list of hand-crafted statistics. The remaining gap suggests that human judgments depend on richer interactions among luminance, structure, texture, and scene content.

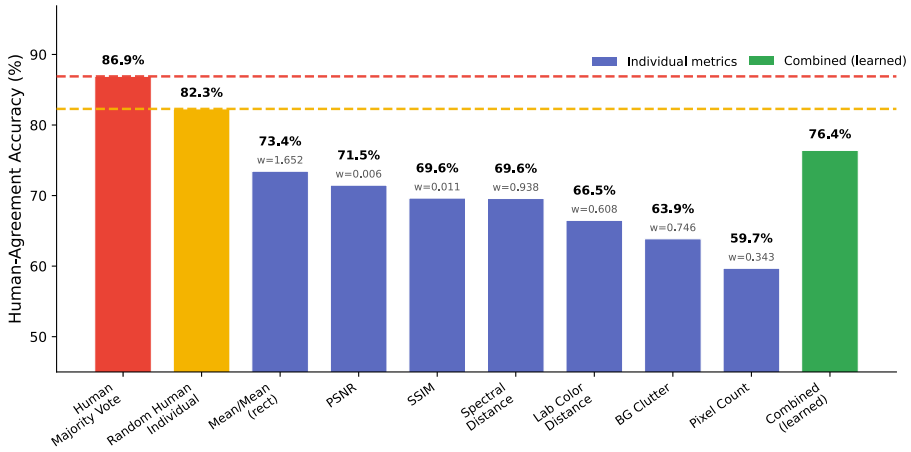

Figure 8: Inter-observer-consistency-weighted accuracy for the low-level visibility metrics on the validation split. The two leftmost bars show the human majority-vote ceiling and the expected performance of a randomly sampled individual observer. Blue bars show the individual hand-crafted metrics, ordered from highest to lowest accuracy, and the green bar shows the learned linear combination. Numbers above the blue bars indicate the learned weights each feature contributes to the combined metric.

C Model and Training Details

C.1 Selected backbone layers

Table 4 lists the representative layers used for each backbone when constructing the deep visibility metric.

Table 4: Representative layers used for each backbone.

Backbone	Early	Mid	Mid-high	CLS / output
VGG-19 BN	features.5	features.25	features.38	classifier output
ResNet-50	layer1	layer2	layer3	classifier output
ResNet-101	layer1	layer2	layer3	classifier output
DINOv2 ViT-S/14	blocks.2	blocks.5	blocks.8	CLS token
DINOv2 ViT-L/14	blocks.5	blocks.11	blocks.17	CLS token
CLIP ViT-L/14	encoder.layers.5	encoder.layers.11	encoder.layers.17	CLS token
DETR (R50)	backbone.layer2	backbone.layer4	encoder.layer2	encoder output

C.2 Training sweeps and selected learning rates

All models were trained for 50 epochs on the same 80/20 train/validation split over unique 2AFC stimulus pairs. We swept multiple hyperparameters for each adaptation regime and selected the best checkpoint separately for each backbone based on validation performance. Learning rate for the best checkpoint is given in Table 5. Each sweep was run on a single H100 GPU for 2-24 hours, depending on the model architecture.

D Haploscope Validation Under Real Additive Light

To assess whether the web-based protocol preserves perceptual relationships under real additive light, we repeated the same 2AFC experiment on an optical haploscope with beam splitters (n=17,

Table 5: Learning rates for the best-performing checkpoint of each backbone and adaptation regime after hyperparameter sweeping.

Backbone	CW LR	LoRA LR	Full FT LR
VGG-19 BN	0.1	—	5×10^{-5}
ResNet-50	0.1	—	5×10^{-5}
ResNet-101	0.1	—	5×10^{-5}
DINOv2 ViT-S/14	0.1	5×10^{-4}	5×10^{-6}
DINOv2 ViT-L/14	0.1	5×10^{-4}	5×10^{-6}
CLIP ViT-L/14	0.1	5×10^{-4}	1×10^{-5}
DETR (R50)	0.1	—	1×10^{-5}

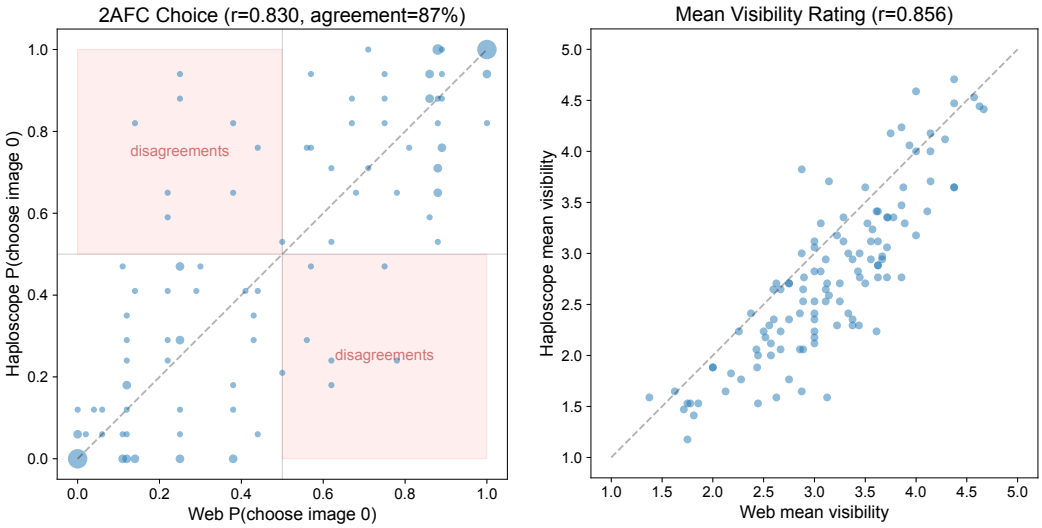

Figure 9: Comparison between the conventional-display and haploscope settings on overlapping stimulus pairs. **Left:** agreement of 2AFC choices across settings. For overlapping points, circle area indicates counts. **Right:** agreement of mean subjective visibility ratings after mapping the five rating categories to integers from 1 to 5.

each completing 120 trials). This setup optically superimposes emitted foreground light onto a real background at the observer's eyes, approximating the image formation process of OST-AR displays. Both the foreground and background were presented at a depth of 1.0 m. The trials were randomly chosen from the web-based study, matching the foreground, background, and their relative luminance.

The results show strong agreement between the two settings, with their majority vote agreed on 87% of trials. See Figure 9 (left) for their respective inter-subject consistency on the same trials, where the circle size represents trial counts. Subjective visibility ratings were also highly correlated across settings (Figure 9, right). For this analysis, we converted the five rating categories to integer values from 1 to 5 as an approximate numeric scale. The web-based ratings were systematically more positive, but the overall ranking across stimuli remained closely aligned.

Predicting haploscope judgments. We evaluated the models trained on web-based judgments against the haploscope responses, without further training or adaptation. We used the same consistency-weighted accuracy as in the main benchmark and recomputed the human reference levels and the low-level baseline performance on haploscope judgments.

Table 6 shows that adaptation to web judgments substantially improves prediction of haploscope responses. The DINOv2 ViT-L/14 LoRA model achieves 79.1%, compared with a 76.1% individual-observer reference and a 68.6% for the low-level metrics combined. These results

Table 6: Consistency-weighted accuracy (%) on haploscope judgments. Models were evaluated without training on haploscope responses. CW denotes channel-weight learning; Full FT denotes full fine-tuning. Dashes indicate configurations not evaluated.

Backbone	Pretrained	CW	LoRA	Full FT
VGG-19 BN	63.1	67.8	—	79.0
ResNet-50	61.6	69.6	—	79.4
ResNet-101	62.1	69.0	—	76.7
DINOv2 ViT-S/14	54.8	65.4	75.7	74.9
DINOv2 ViT-L/14	52.2	66.3	79.1	76.9
CLIP ViT-L/14	61.7	66.4	78.3	77.7
DETR (R50)	54.4	67.6	—	77.9
Majority reference		82.6		
Individual-observer reference		76.1		
Learned low-level baseline		68.6		

demonstrate strong transfer across viewing conditions and support our web-based protocol as a low-cost proxy for measuring visibility under real additive light.

E Feature Ablations and Low-Level Cue Fusion

For this section, we used the trained network model DINOv2 ViT-S/14 with LoRA as baseline (81.8% accuracy).

E.1 Complementarity with low-level cues

We combined each network layer (or all four) with each low-level cue (or their learned combination). Each combination was $s = az_{\text{low}} + bz_{\text{net}}$, with both scores standardized using means and standard deviations of the training 2AFC trials. We fixed the features and only fitted the two unconstrained weights a and b on the training set. Validation used the same normalization statistics. We measured accuracy gains over each component alone on identical validation trials (Table 7).

Table 7: Changes in consistency-weighted accuracy from cue fusion, in percentage points. **Left:** adding low-level cues to the full network metric, relative to the network metric alone. **Right:** adding network layers to the learned low-level combination, relative to the low-level baseline alone.

Added low-level cue	Change in pp
Luminance ratio	$+0.04$
Lab color distance	$+0.04$
Spectral distance	$+0.05$
Background clutter	$+0.14$
Foreground pixel count	$+0.09$
SSIM	-0.07
PSNR	-0.12
Learned combination	-0.02

Added network score	Change in pp
Early	$+0.99$
Mid	$+1.28$
Mid-high	$+2.46$
CLS	$+4.27$
All four levels	$+5.36$

Low-level cues improve the full network's accuracy by at most 0.14 percentage points. Conversely, network scores improve the learned low-level baseline by up to 5.36 percentage points, with larger gains from deeper features. These results support that our deep metrics already subsume the individual low-level perceptual mechanisms, while those mechanisms can benefit from high-level scene understanding.

E.2 Contributions of individual feature levels

We computed a visibility score from each network layer alone and from all levels except one (Table 8).

Table 8: Layer ablations and the resulting consistency-weighted accuracy (%). The full metric combines all four feature levels.

Feature level	Only this level	Excluding this level
Early	75.9	81.8
Mid	76.9	81.7
Mid-high	77.8	81.5
CLS	79.2	77.3
All four	81.8	—

Removing the final CLS contribution reduces accuracy by 4.5 percentage points, whereas removing any one of the three intermediate feature levels changes accuracy by at most 0.3 percentage points. The CLS representation thus makes the largest individual contribution in this ablation, while the three early/intermediate layers carry redundant information and may be removable under resource constraints.

F Correlation Between Human and Model Matched Luminance

In addition to mean absolute error, we evaluated each predictor by the Spearman correlation between the model-predicted matched luminance and that of the human group median. Figure 10 shows the same qualitative pattern as the MAE analysis in the main text: off-the-shelf pretrained features and the low-level baseline trail behind the learned models, while training with 2AFC supervision consistently improves correlation across backbone families and adaptation regimes.

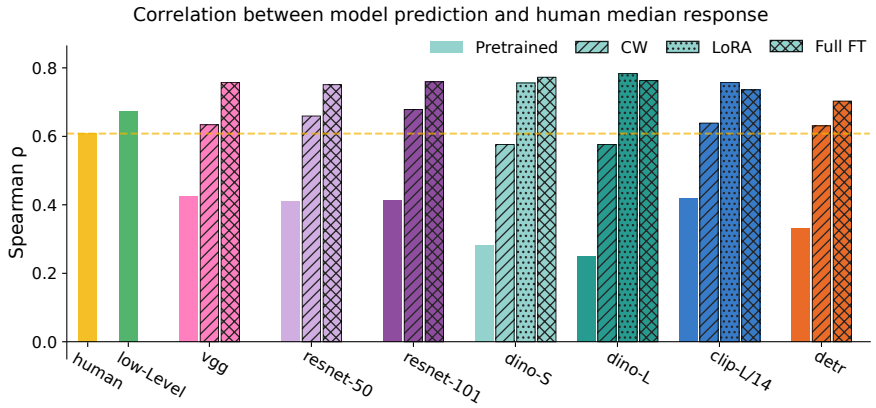

Figure 10: Spearman correlation between model-predicted matched luminance and human group median. Bars report the human leave-one-out reference, the combined low-level baseline, and each model under pretrained, channel-weight (CW), LoRA, and full fine-tuning regimes.

G Content-placement example

Figure 11 illustrates one example application of our visibility metric: comparing candidate placements of the same foreground within a scene. The heatmap shows where the model predicts greater visibility for the person to be placed.

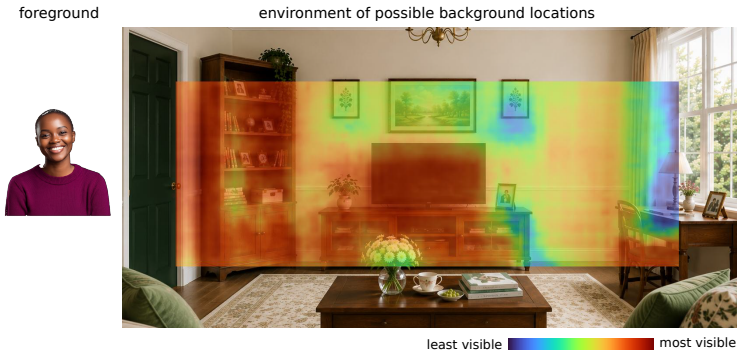

Figure 11: **Deep-ARViM for content placement. Left:** True foreground size. **Right:** Our metric produces visibility heatmaps that identify where the same AR foreground would be most visible within a scene. Note that the window on the right makes that side of the scene brighter.

H Computational Demand

One important finding of our work is that smaller models such as ResNet-50 can also have competitive performance on visibility prediction, if trained on the proper data and task. Although the heaviest model is not always required, here we argue even the heavier configuration of our metric can fit the compute budget of currently shipping smart glasses, under the plain implementation that can potentially be further optimized.

H.1 Memory estimates

We believe memory may not be a concern for our metric because it is very likely for some pre-trained vision models to already exist on device for other purposes, such as for scene recognition. Our results suggest that applying a light LoRA adaptor on these models is sufficient for visibility prediction, at minimal additional cost.

Despite that, we report here the full memory footprint needed for our metric under fp16 and at 224×224 pixels, except DETR at 600×600 pixels (Tables 9).

The memory estimate includes backbone weights, retained feature outputs for both the background and the composite ("Taps"), and transient forward-pass activations ("Act").

Table 9: Approximate fp16 memory requirements in MB.

Backbone	Resolution	Weights	Taps ($\times 2$)	Act	Total
VGG-19	224^2	274.0	16.85	12.2	303.1
ResNet-50	224^2	48.7	5.36	3.1	57.2
ResNet-101	224^2	85.0	5.36	3.1	93.4
DINO-S	224^2	42.1	1.13	1.5	44.7
DINO-L	224^2	580.5	3.02	4.0	587.6
CLIP-L/14	224^2	578.3	3.02	4.0	585.3
DETR-R50	600^2	79.2	14.51	22.0	115.7

H.2 Compute and latency

Table 10 reports the number of multiply-accumulate operations (in GMACs) per backbone forward pass and a compute-based latency estimate for two passes, one for the background and one for the composite. The numbers are estimated based on the Qualcomm Snapdragon AR1 Gen 1. It is the system-on-chip in the Meta Ray-Ban Display (released late 2025), one of the consumer smart-glasses whose information we can find online. This is an illustrative throughput assumption,

not a measured rate on an AR device. We found that inference speed is well beyond what visibility assessment needs, which operates as the scene changes.

Table 10: Approximate backbone compute and projected fp16 latency.

Backbone	Resolution	GMACs / pass	Latency, two passes (s)
VGG-19	224^2	19.63	0.196
ResNet-50	224^2	4.09	0.041
ResNet-101	224^2	7.80	0.078
DINO-S	224^2	6.12	0.061
DINO-L	224^2	81.01	0.810
CLIP-L/14	224^2	81.01	0.810
DETR-R50	600^2	34.48	0.345

I Full Instructions Given to the Participants

This section provide screenshots that include full instructions given to the participants at experiment time. Subjects need to choose "Yes" to all questions before proceeding with the experiment.

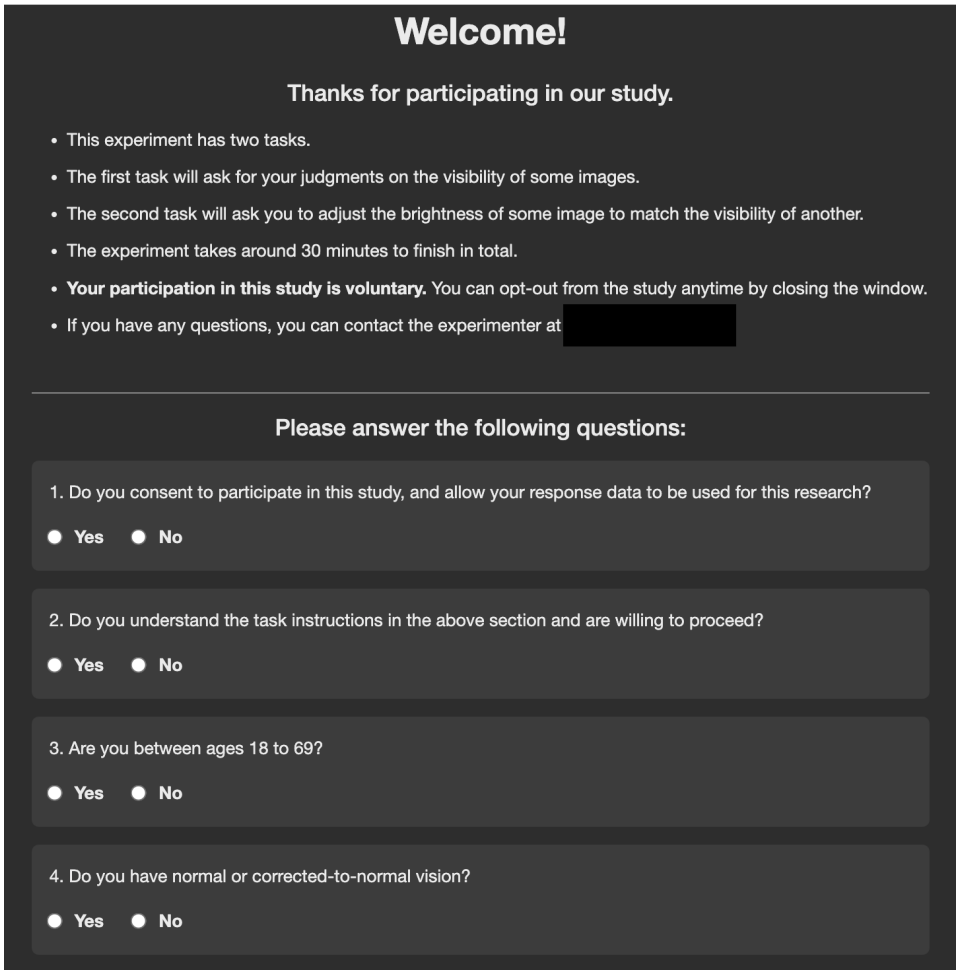

Figure 12: Experiment start page.