What do saliency models predict?

based on: Koehler, Guo, Zhang, & Eckstein (JoV, 2014)

Zoya Bylinskii, March 31, 2014 - paper presentation for 9.S912

Are human eye movements better predicted by bottom-up or top-down information?

- One view: when free-viewing images, people are drawn to conspicuous or salient regions (areas/objects that stand out against the background)
- Alternative view:
  - Bottom-up models do not fully account for human eye movements
  - Search target information and context offer important top-down information for directing eye movements
  - Fixations are often directed to objects
- Debate remains about importance + contribution of bottom-up information
- Task constraints may determine the contributions of top-down and bottom-up processes
- Computational saliency models take as input an image and return a topographical map of salient image locations
  - can predict for basic, bottom-up influences (Foulsham & Underwood, 2008)

Itti-Koch (IK) Model

- Original model proposed by Koch & Ullman (1985), updated by Itti & Koch (2000), Walter & Koch (2006)
- Combination of weighted feature maps, WTA mechanism

Judd et al., MIT Tech Report 2012

Attention based on information maximization (AIM) Model

- Bruce & Tsotsos (2009)
- use of global context, pyramidal architecture, inhibition, feature fusion

Judd et al., MIT Tech Report 2012

Saliency using natural statistics (SUN) Model

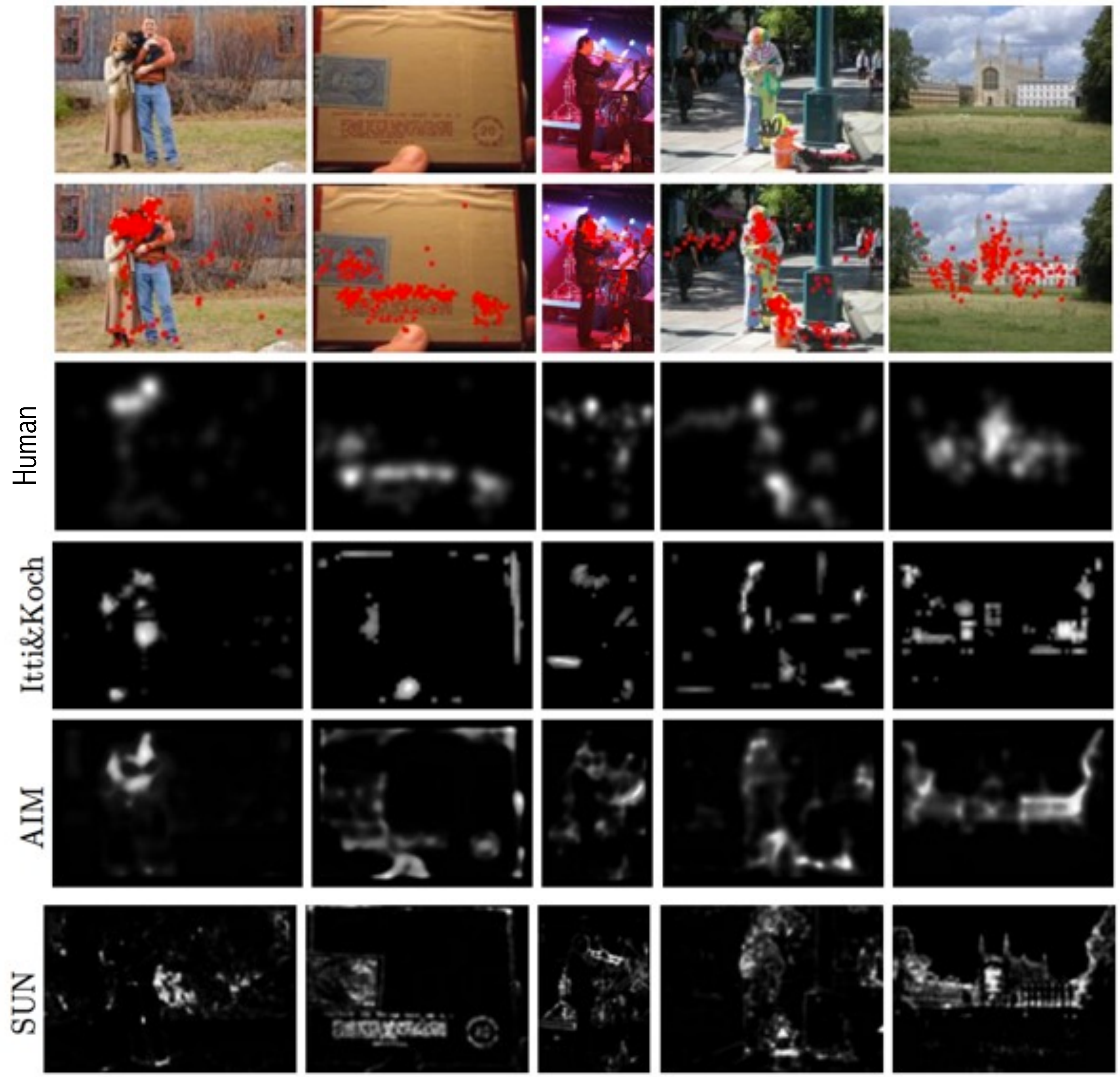
- Zhang et al. (2008)
- "amount of surprise" in each image region
- point-wise mutual information of each point in an image (Bayesian framework)

Judd et al., MIT Tech Report 2012

Model Comparison (on Judd's saliency benchmark dataset)

Judd et al., MIT Tech Report 2012

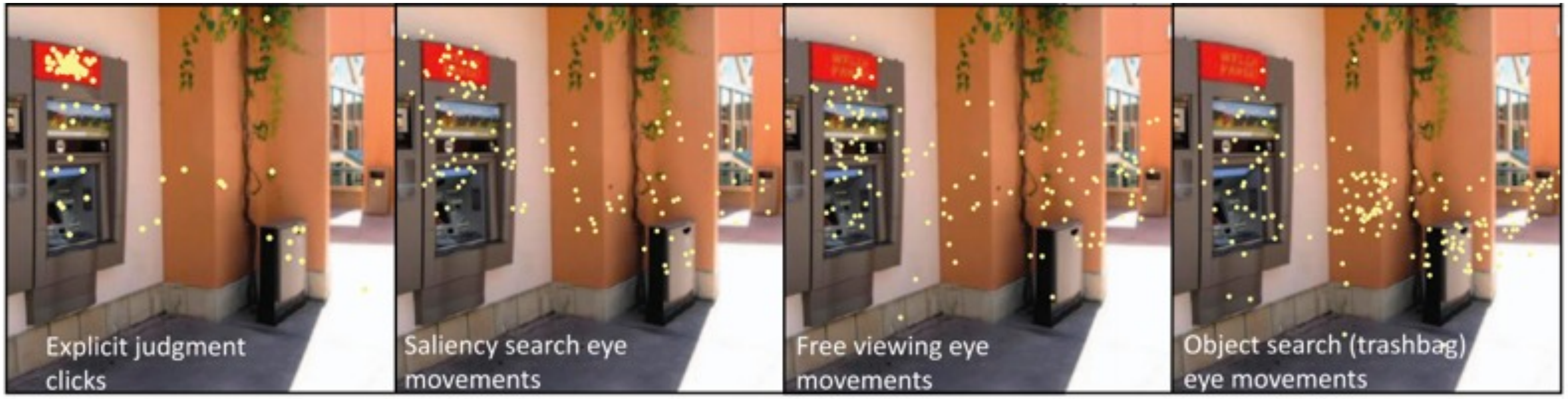
Model Comparison (on Judd's saliency benchmark dataset)



Maps have been histogram-normalized to facilitate visual comparison

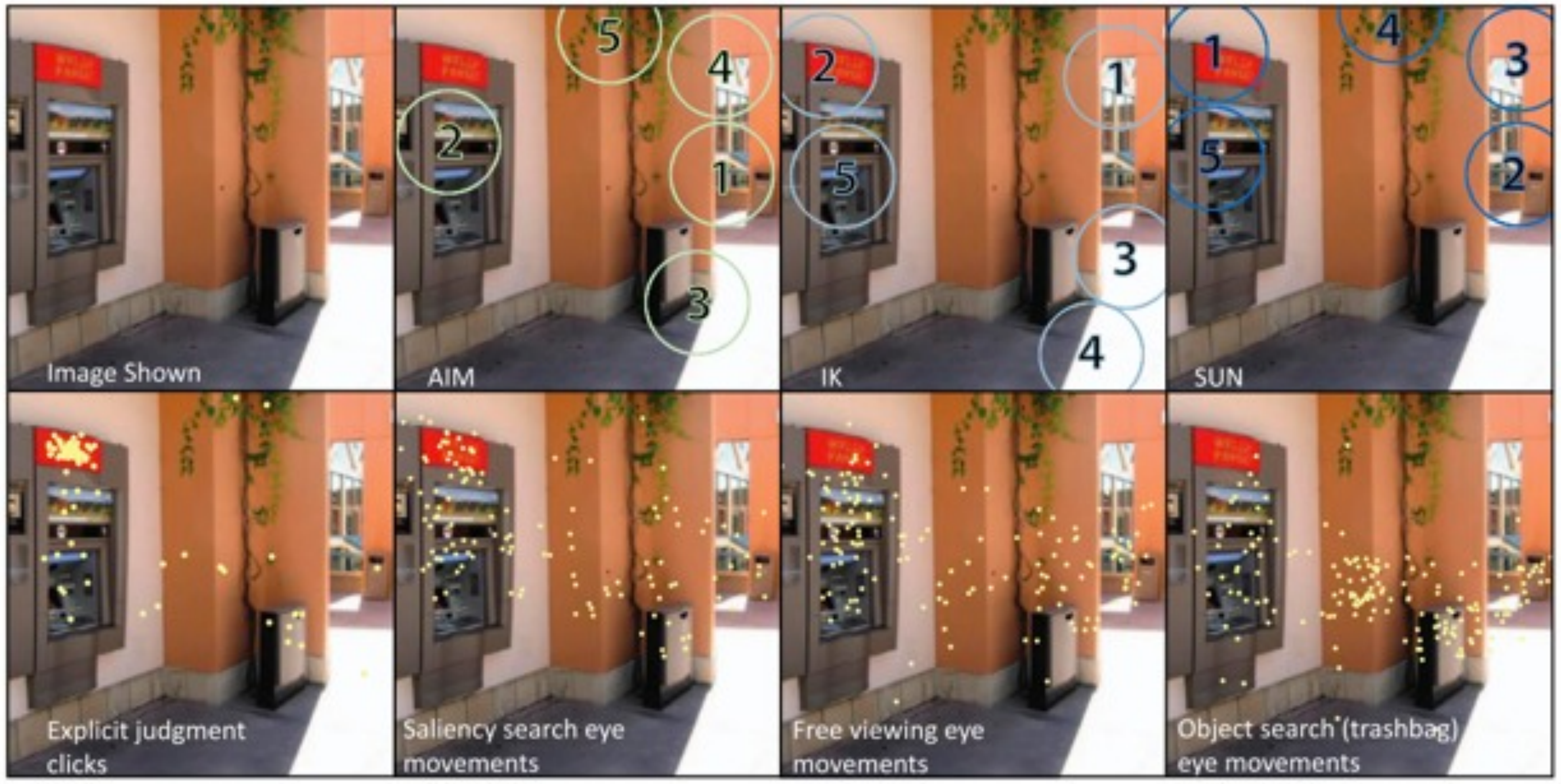
Judd et al., MIT Tech Report 2012

Tasks



- Explicit judgement: click on object/area that is most salient ("something that stands out or catches your eye")
- Free viewing: free view the image (no other instructions)
- Saliency search: determine whether the most salient object/location is on the left or right side of the image
- Cued object search: determine whether or not a target object is present in the image (cued with object name)
- 800 images shown (400 for the last task), for 2 sec each for each of the last 3 (eyetracking) tasks

Maps and fixation controls



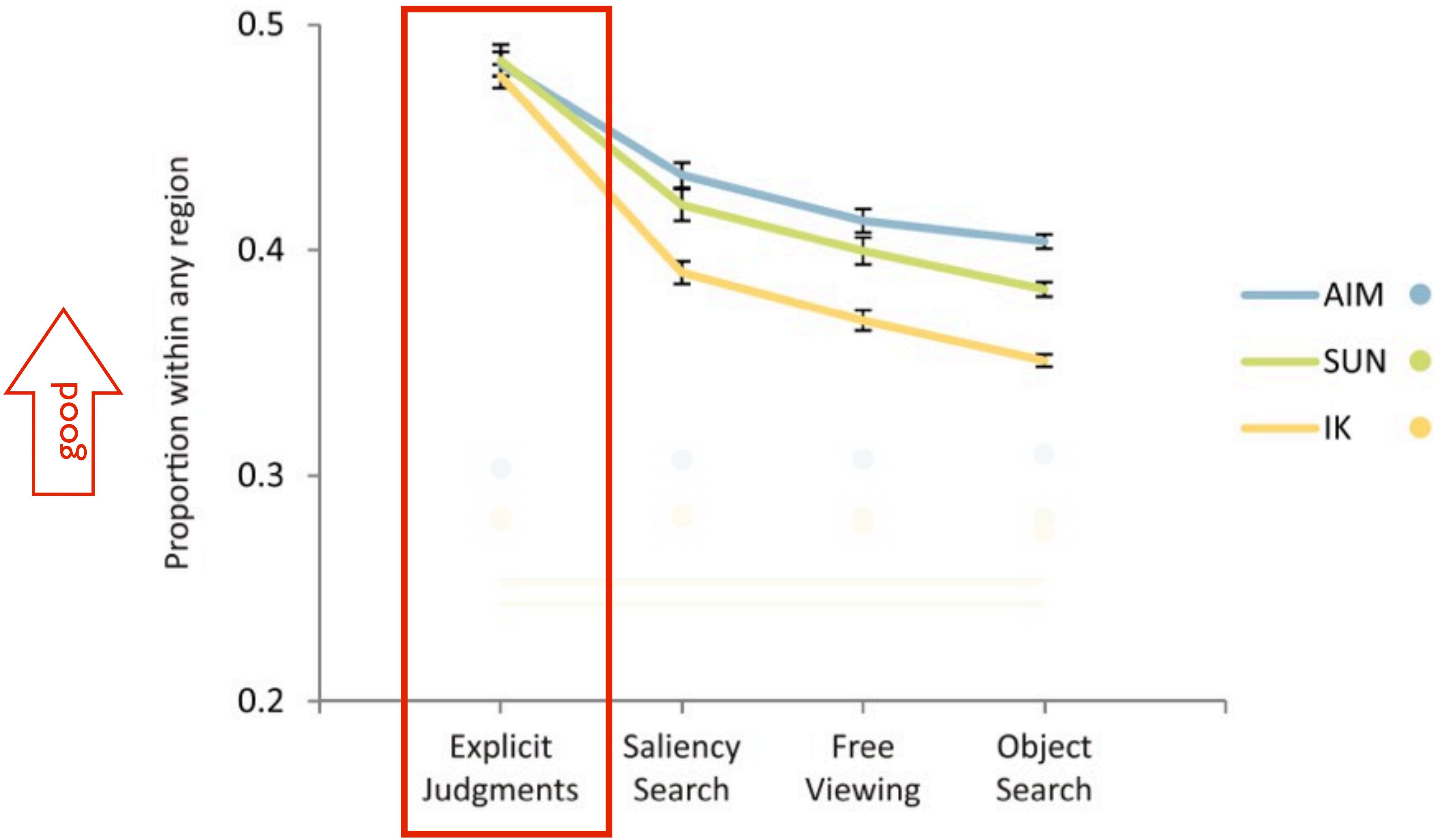
approx. size of human fovea

- Saliency maps computed using 3 models for all 800 images, original settings
- Custom algorithm used to select top 5 non-overlapping salient regions (circular regions with 2 deg. of visual angle)
- Control 1: randomly sampled fixations (uniform probability across all image pixels)
- Control 2: fixations from a different, randomly assigned images (captures observer bias, center bias)

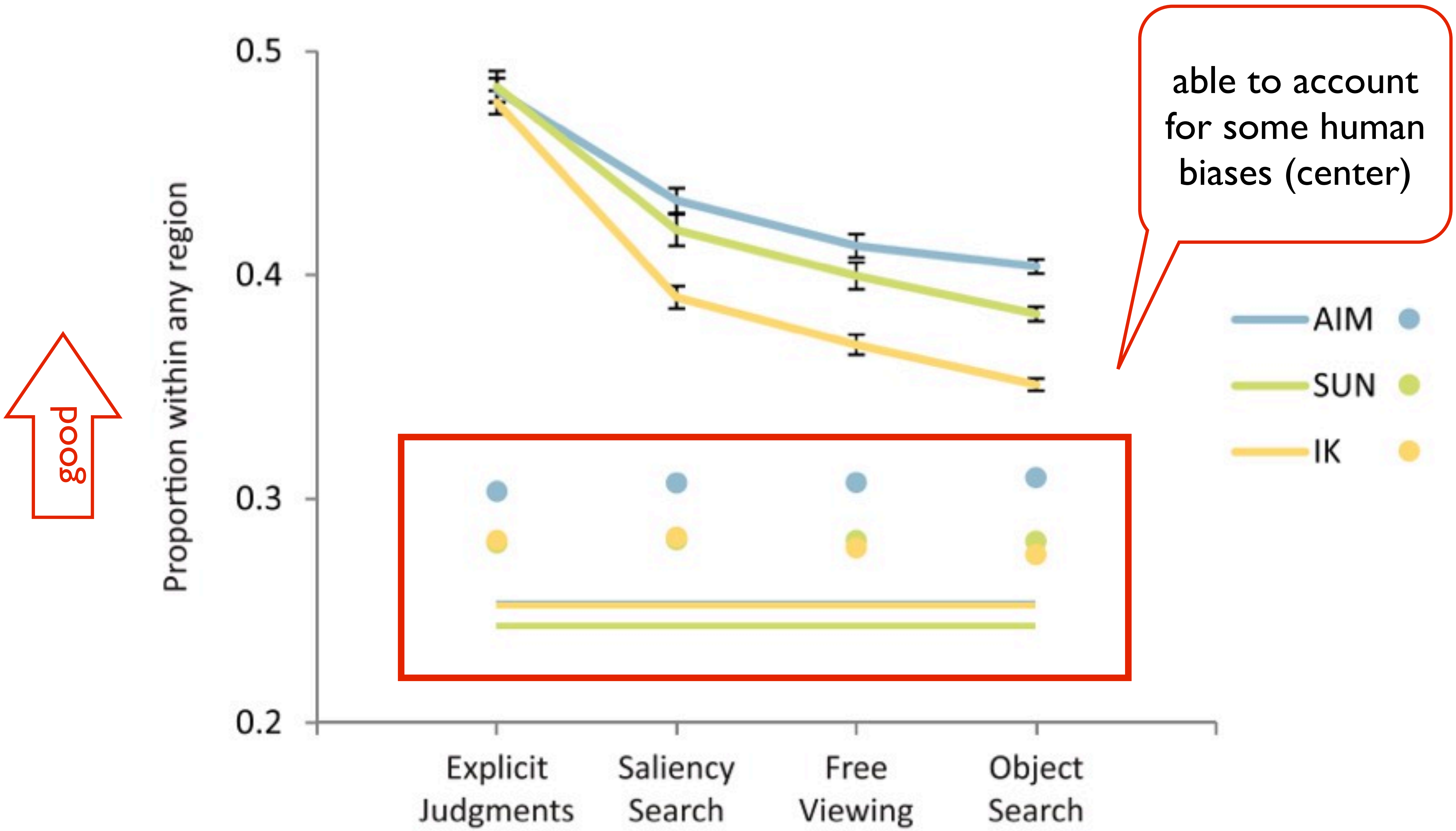
Performance Metric 1: ROI

- proportion of clicks/fixations within top 5 regions of each model (within each, and averaged across, regions)
- measures how well the model predictions are encompassing human behavior

ROI metric shows:
models are better at predicting explicit judgment clicks than any of the fixation tasks

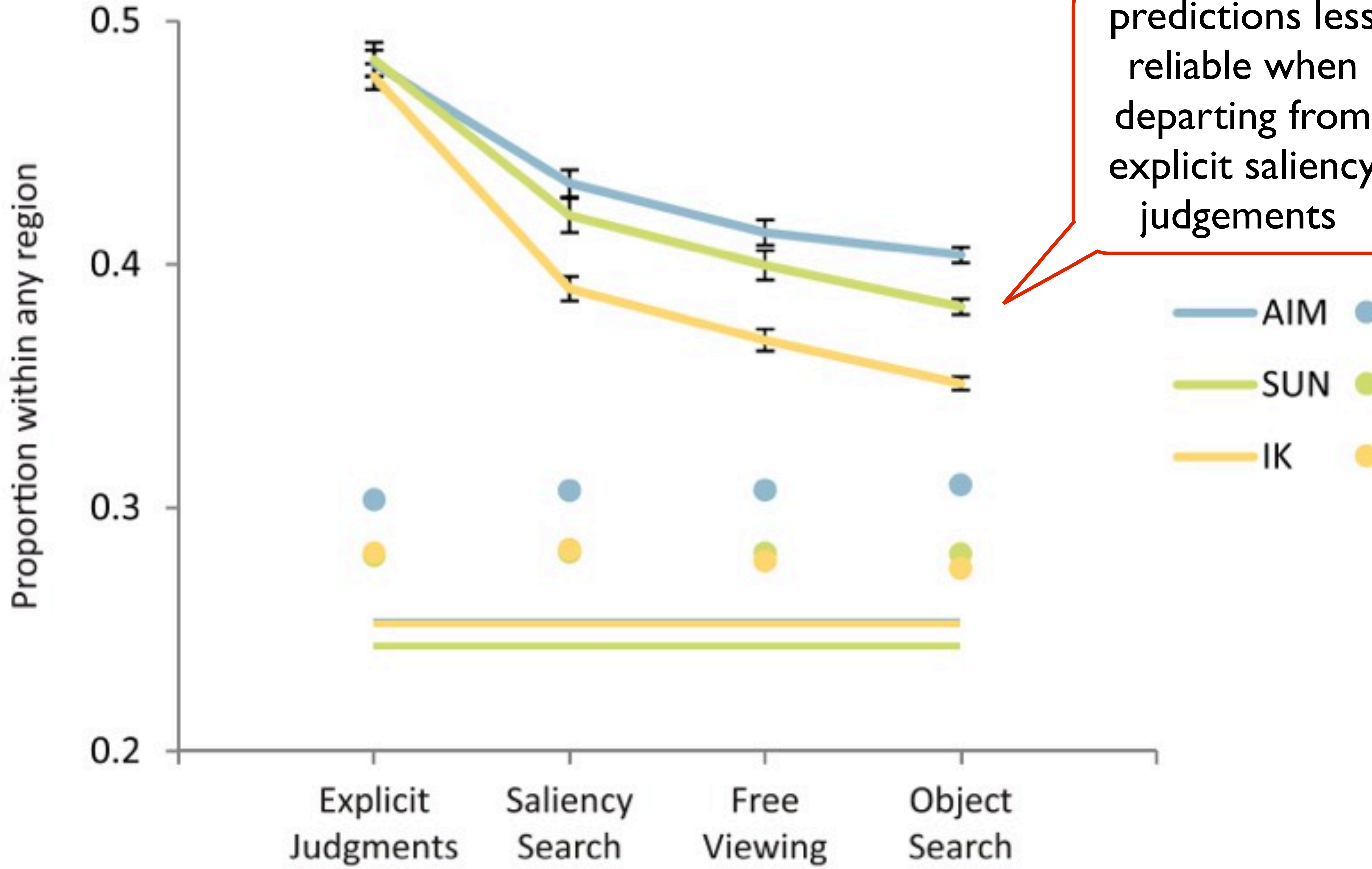


ROI metric shows:
models are better at predicting permuted clicks than random clicks



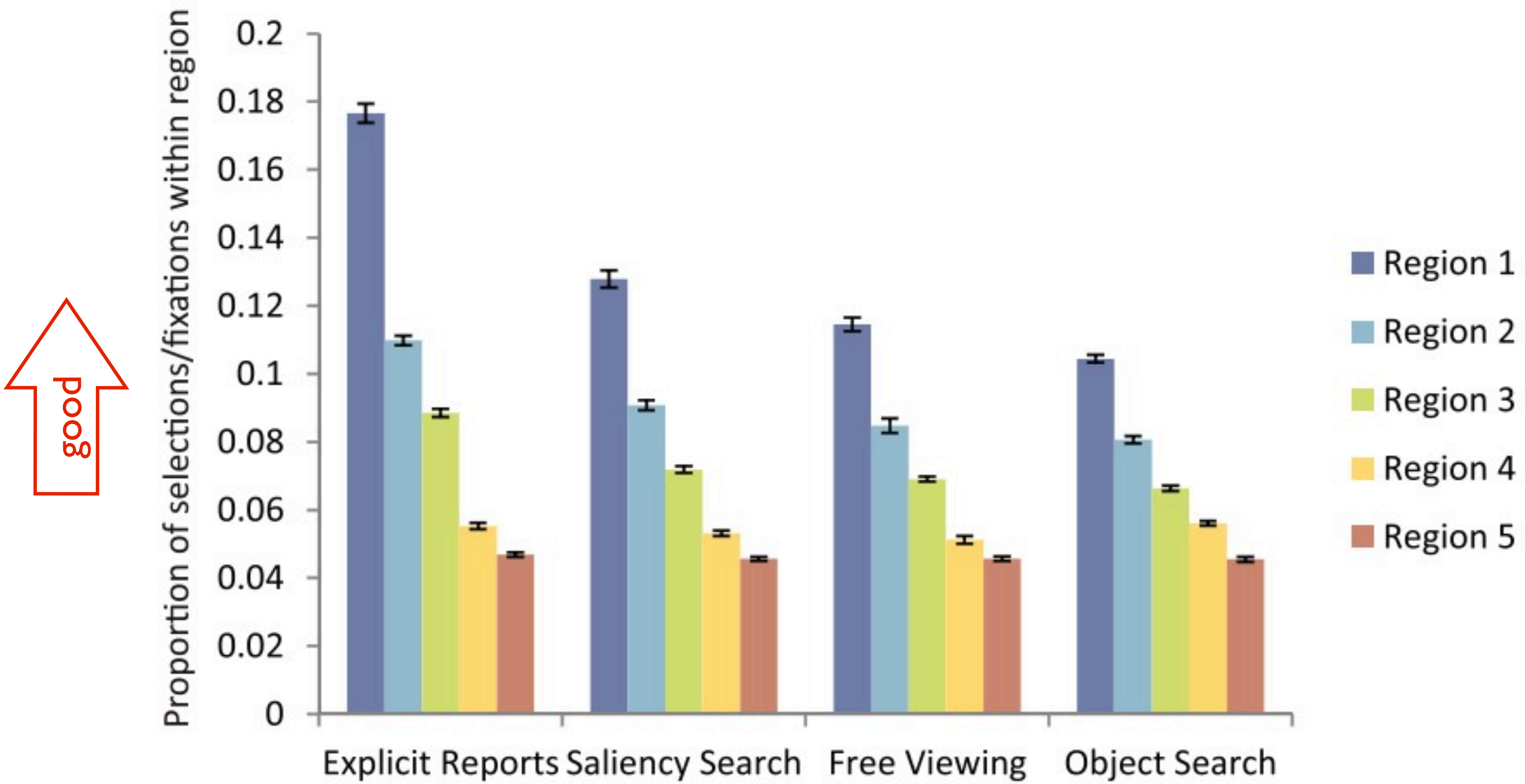
ROI metric shows:
AIM model best able to predict human behavior, across all tasks
SUN model better than IK model, across all tasks
IK shows highest degradation across tasks
AIM shows lowest degradation across tasks

good



predictions less reliable when departing from explicit saliency judgements

ROI metric shows:
higher saliency values in maps are more likely to predict human fixation locations across all tasks

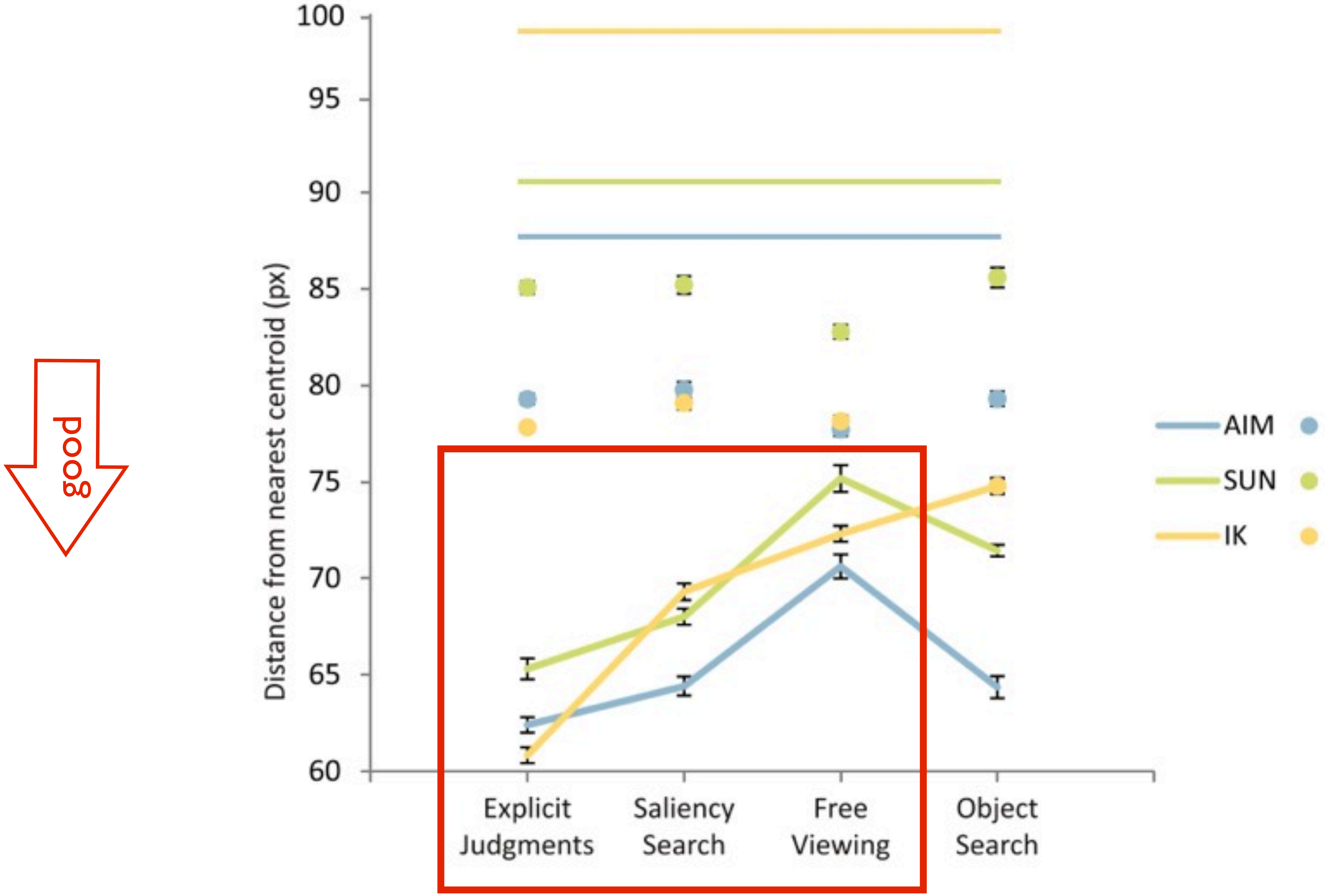


graph for IK model (similar trends for AIM and SUN)

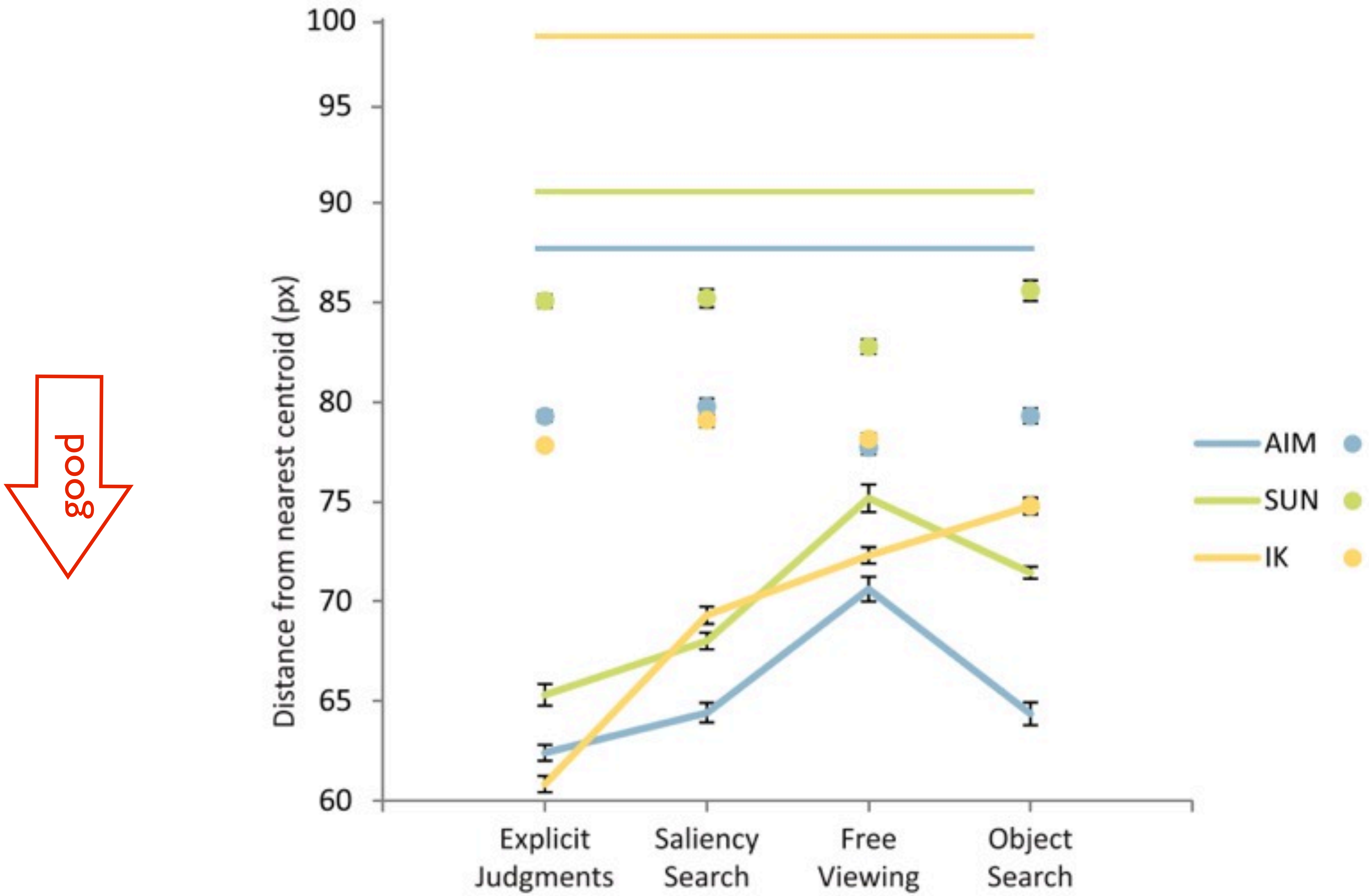
Performance Metric 2: Distance

- average distance of click/fixation to nearest of top 5 regions of each model (continuous measurement)
- measures how tightly clustered a group of fixations are around a salient location

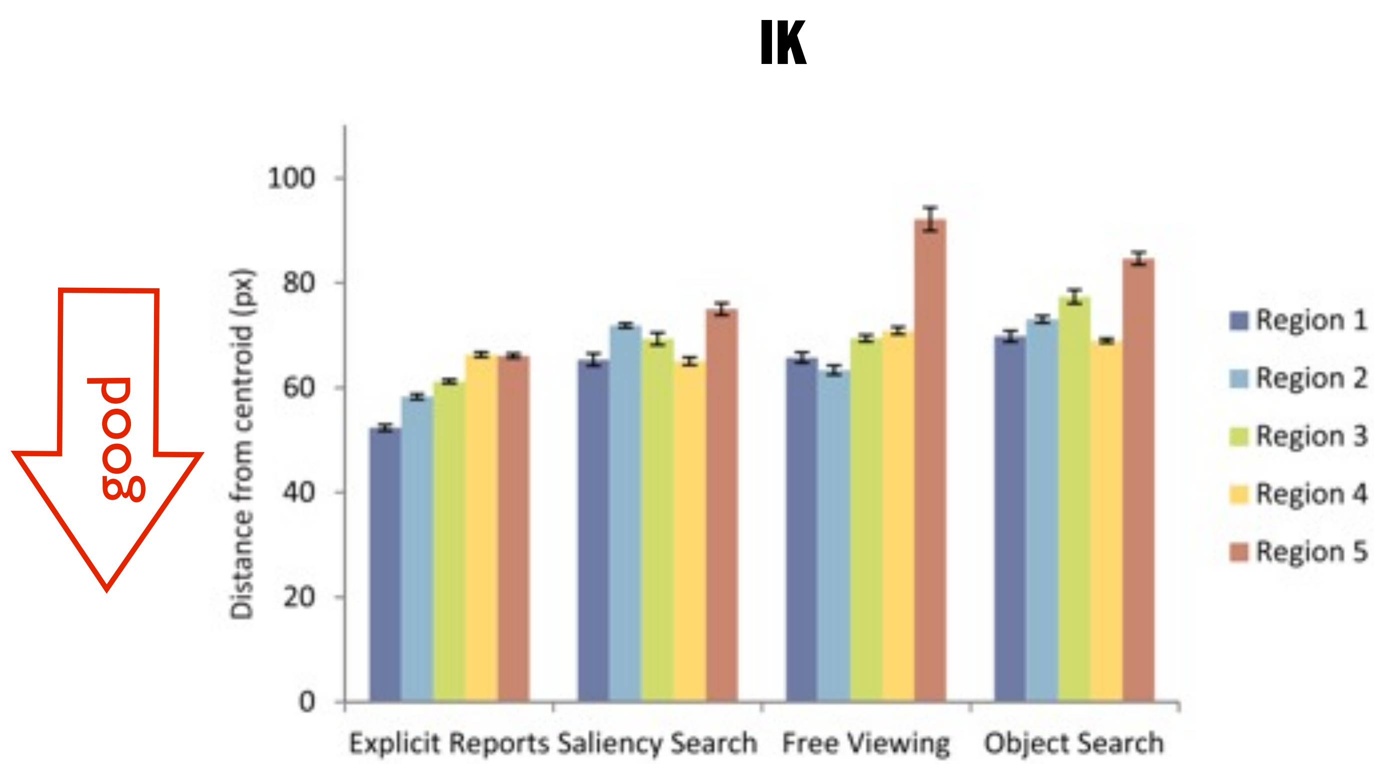
Distance metric confirms:
models are better at predicting explicit judgment clicks than any of the fixation tasks
and shows:
models are better at predicting saliency search than free viewing

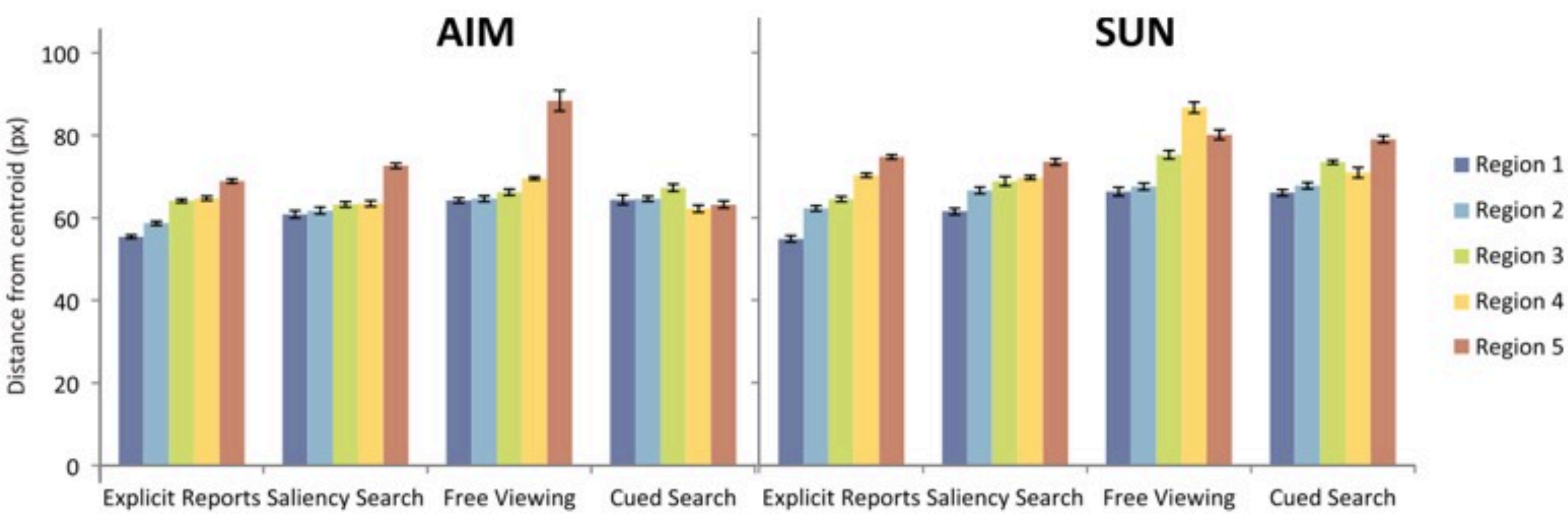


Distance metric confirms:
AIM model best able to predict human behavior, across all tasks
and shows:
IK model better than SUN model at predicting human behavior



Distance metric confirms:
highest saliency values more closely align with human behavior

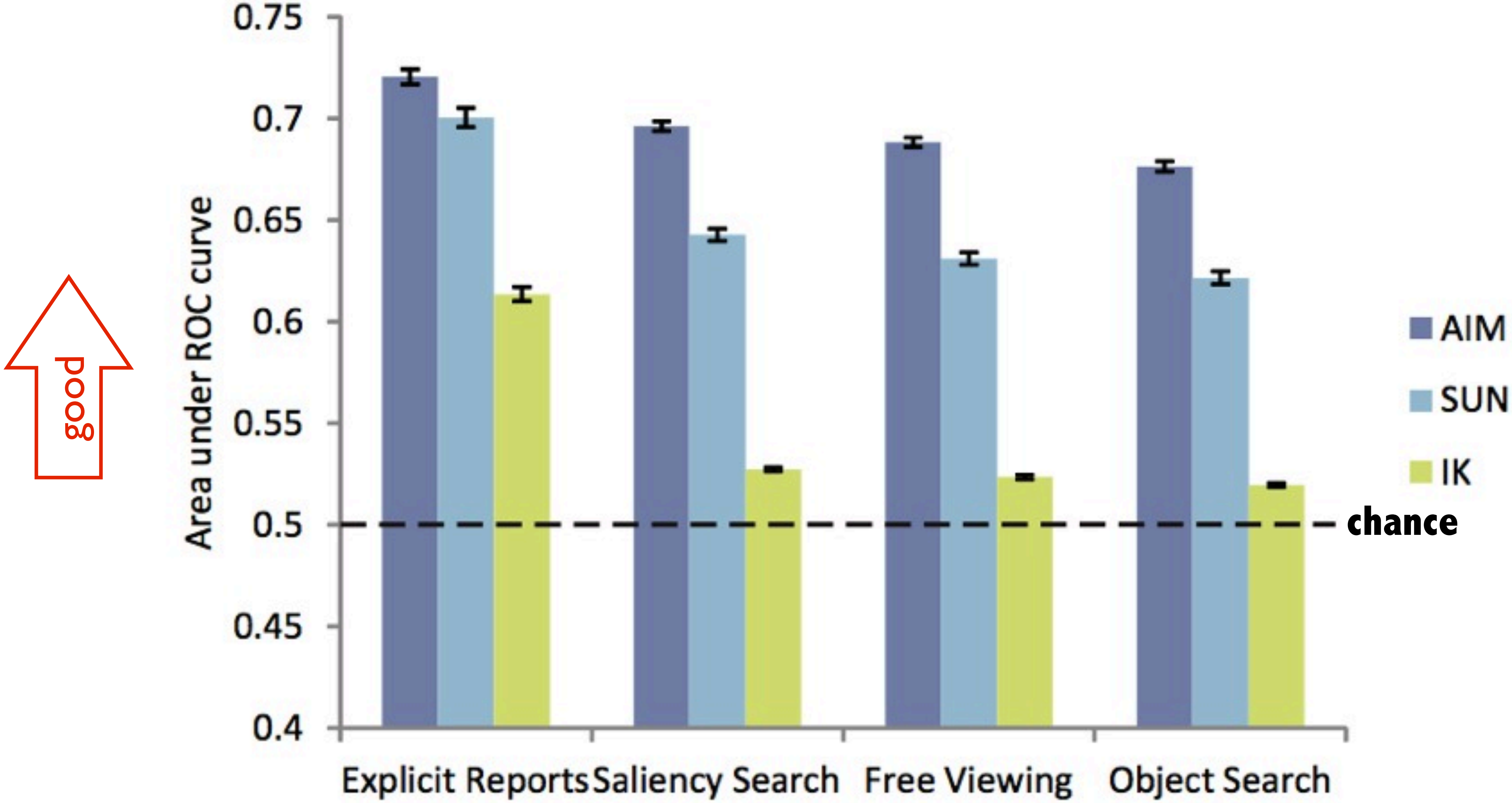




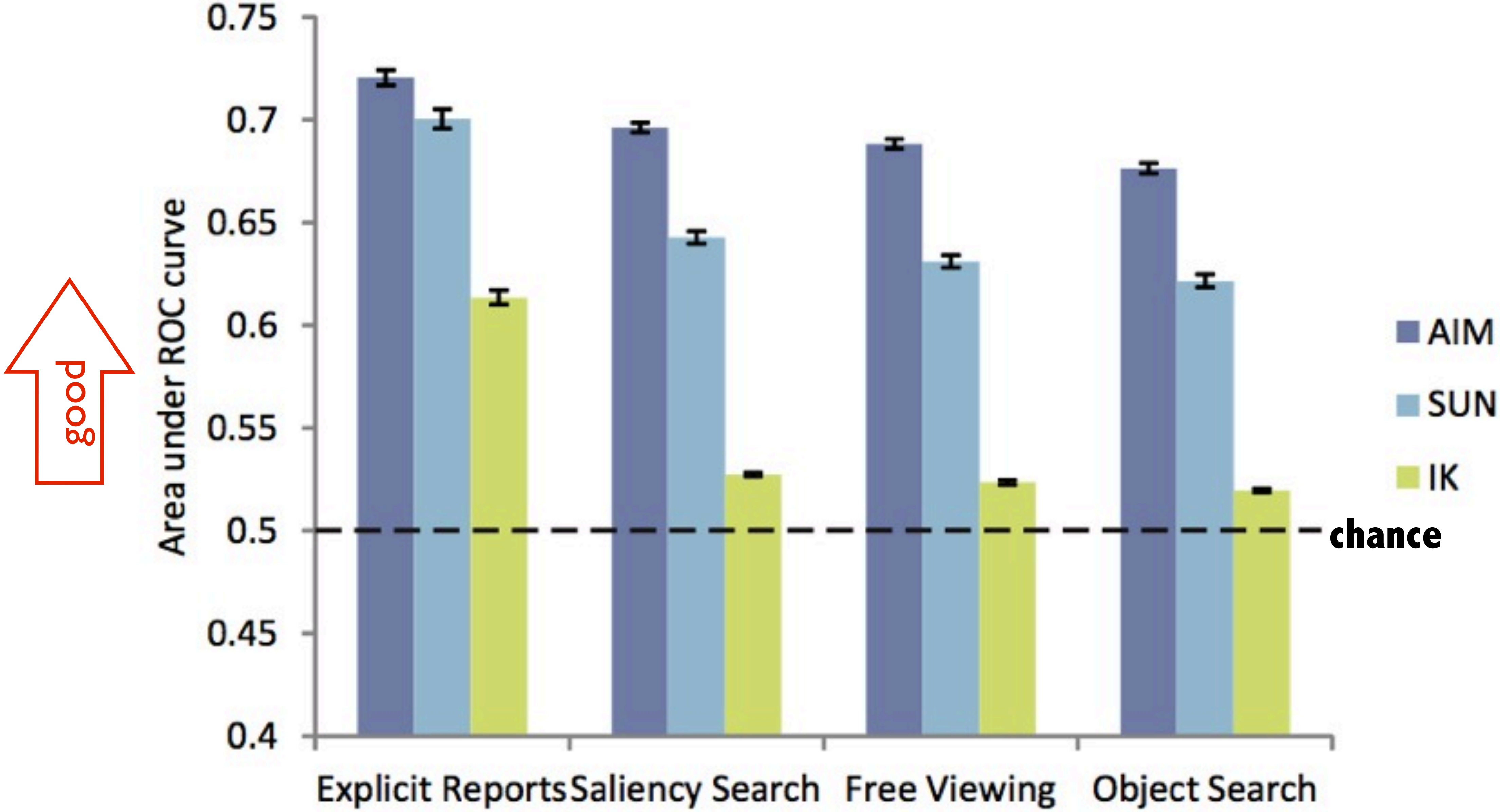
Performance Metric 3: ROC

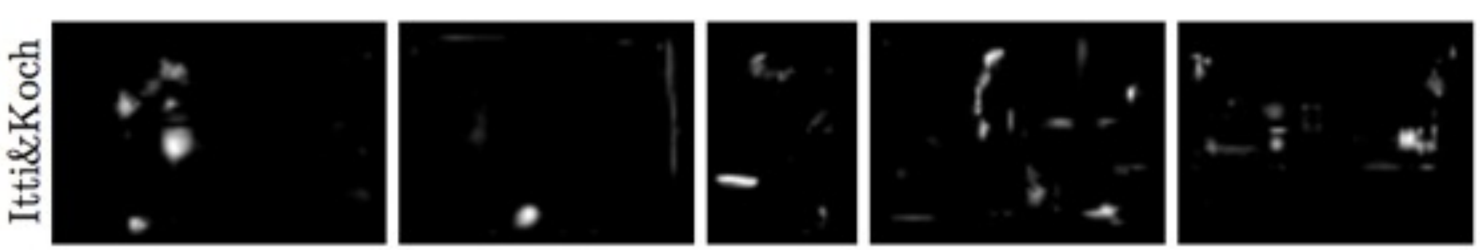
- saliency maps binarized at thresholds varying by 0.001 between 0 and 1 and compared to binary human click/fixation maps, number of hits + false alarms tallied, and point plotted on ROC curve for each threshold
- relatively stable across small image variances, model parameters

ROC metric confirms:
models are better at predicting explicit judgment clicks than any of the fixation tasks
and shows:
models are better at predicting saliency search than object search
models are better at predicting free viewing than object search



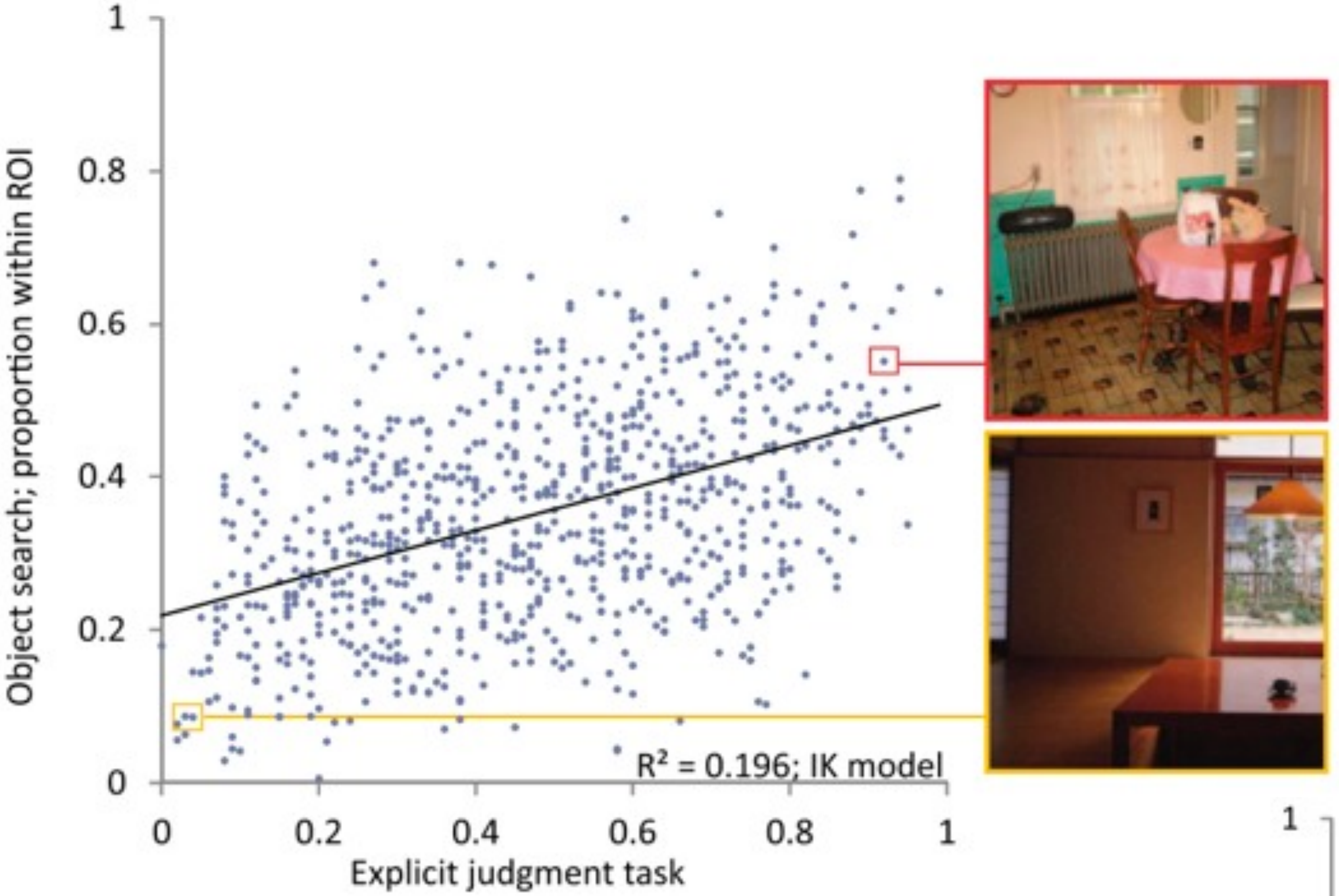
ROC metric confirms:
AIM model best able to predict human behavior, across all tasks
and shows:
SUN model better than IK model





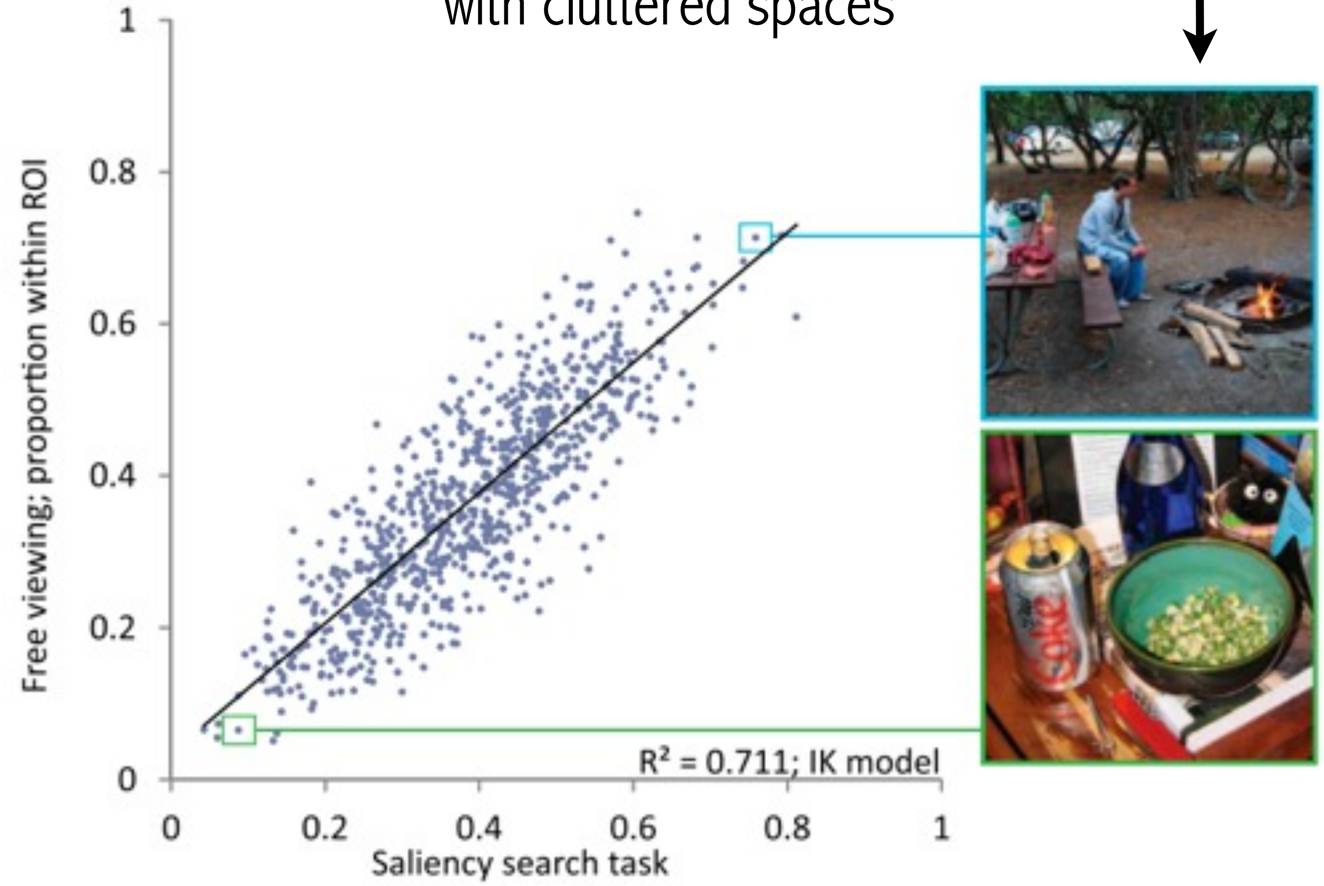
IK model performs poorly on ROC analysis due to its sparse output

Correlation of model metrics across tasks



models are good predictors of human behavior for images with few or stand-out objects

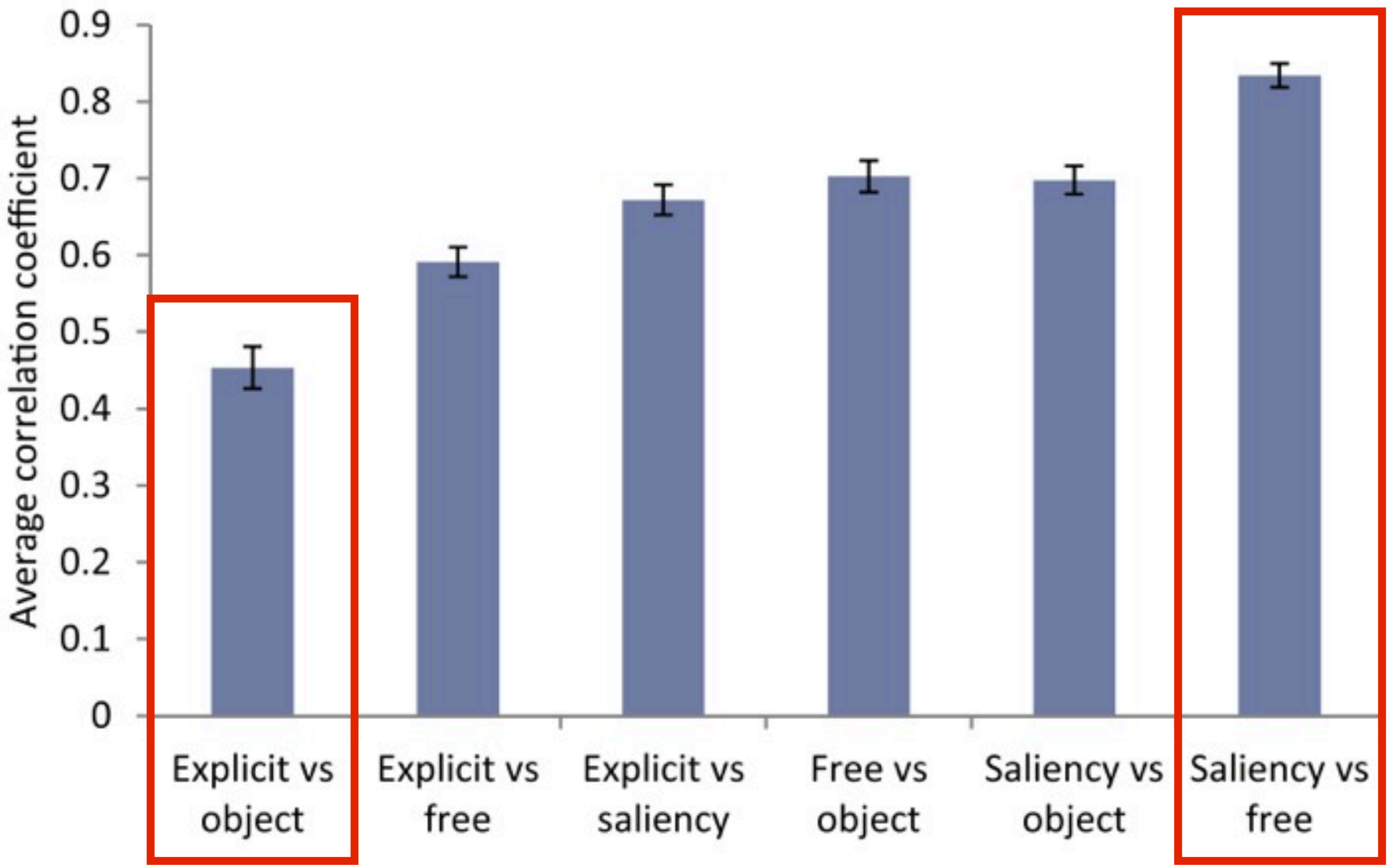
models poorly predict human behavior for images without any distinct salient objects or with cluttered spaces



2 representative scatter plots shown here
(for IK model, with ROI metric)

In total: 54 correlations ran
(3 models x 3 metrics x 6 task pairs)

Correlation of model metrics across tasks shows:
strongest correlation between saliency search and free viewing
weakest correlation between explicit judgement and object search

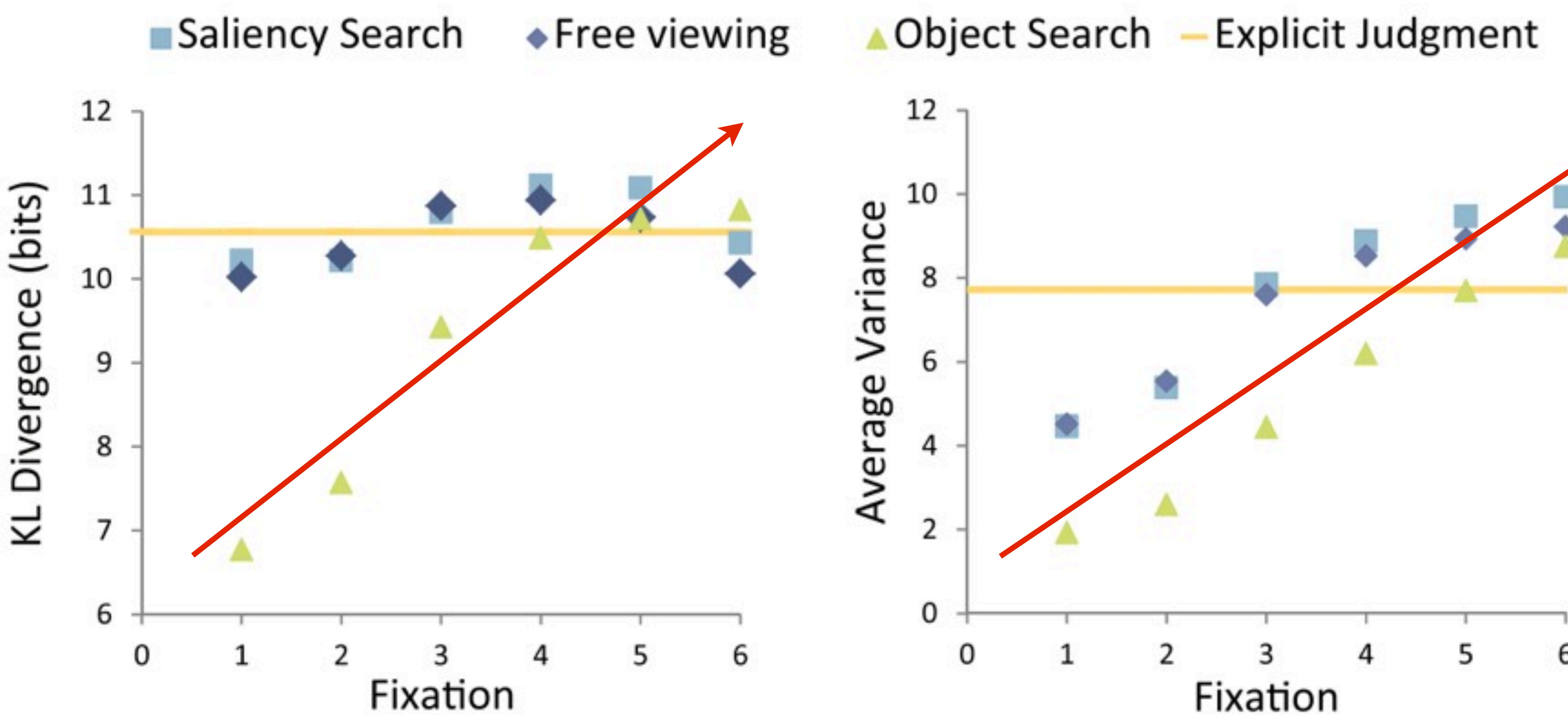


correlations have been averaged across models and metrics

Measuring Individual Differences

- Variance
  - average variance across observers, across x- and y-coordinate fixation locations, across 6 fixations
- Kullback-Leibler divergence
  - KL diverge for each observer against all other observers, averaged across observers + images
  - observer fixation distribution (calculated by diving each image into bins, and tallying fixations per bin)
  - the less similar the distributions, the greater the KL divergence

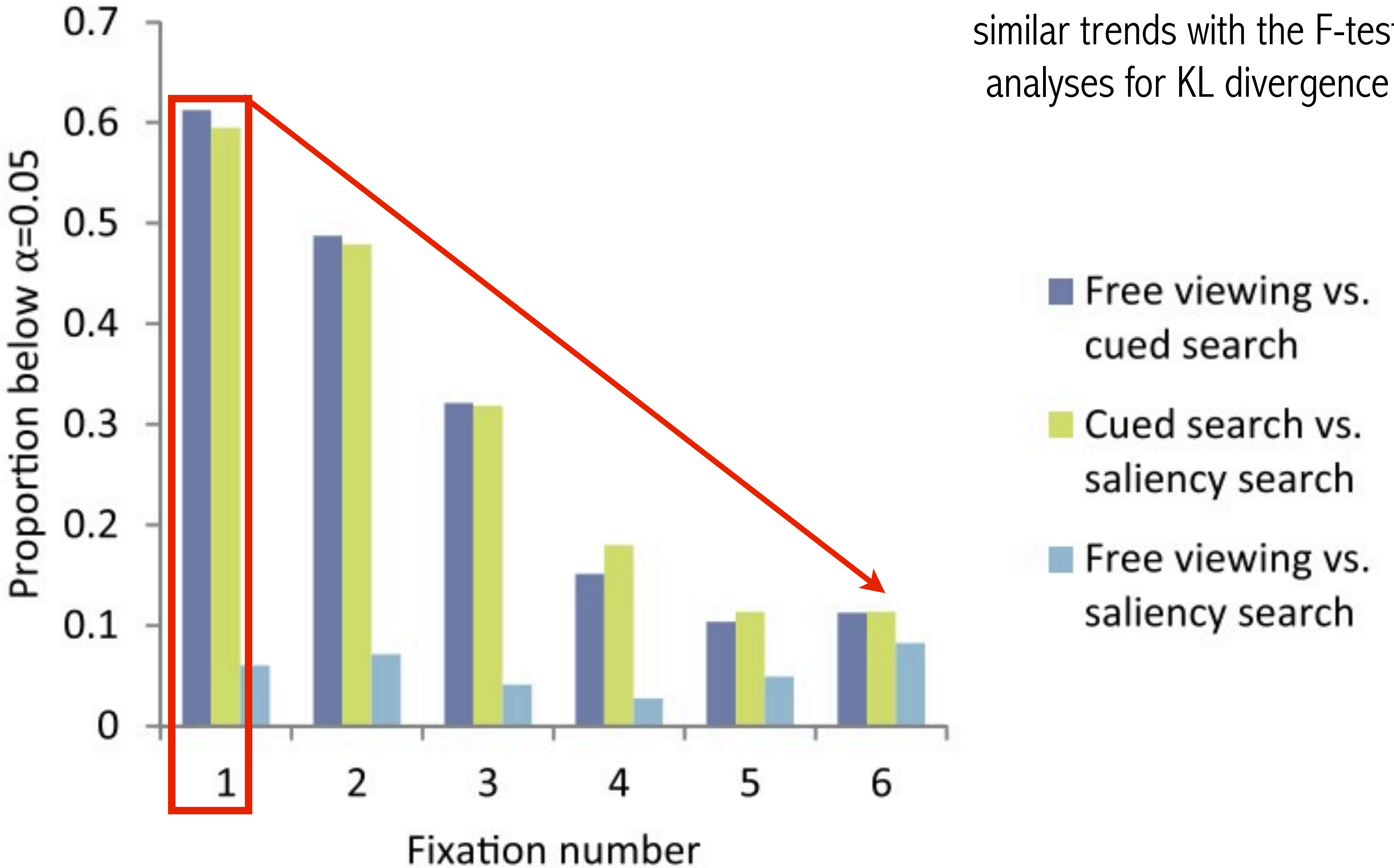
KL divergence and variability in location: increase for later fixations and were lowest for the object search task



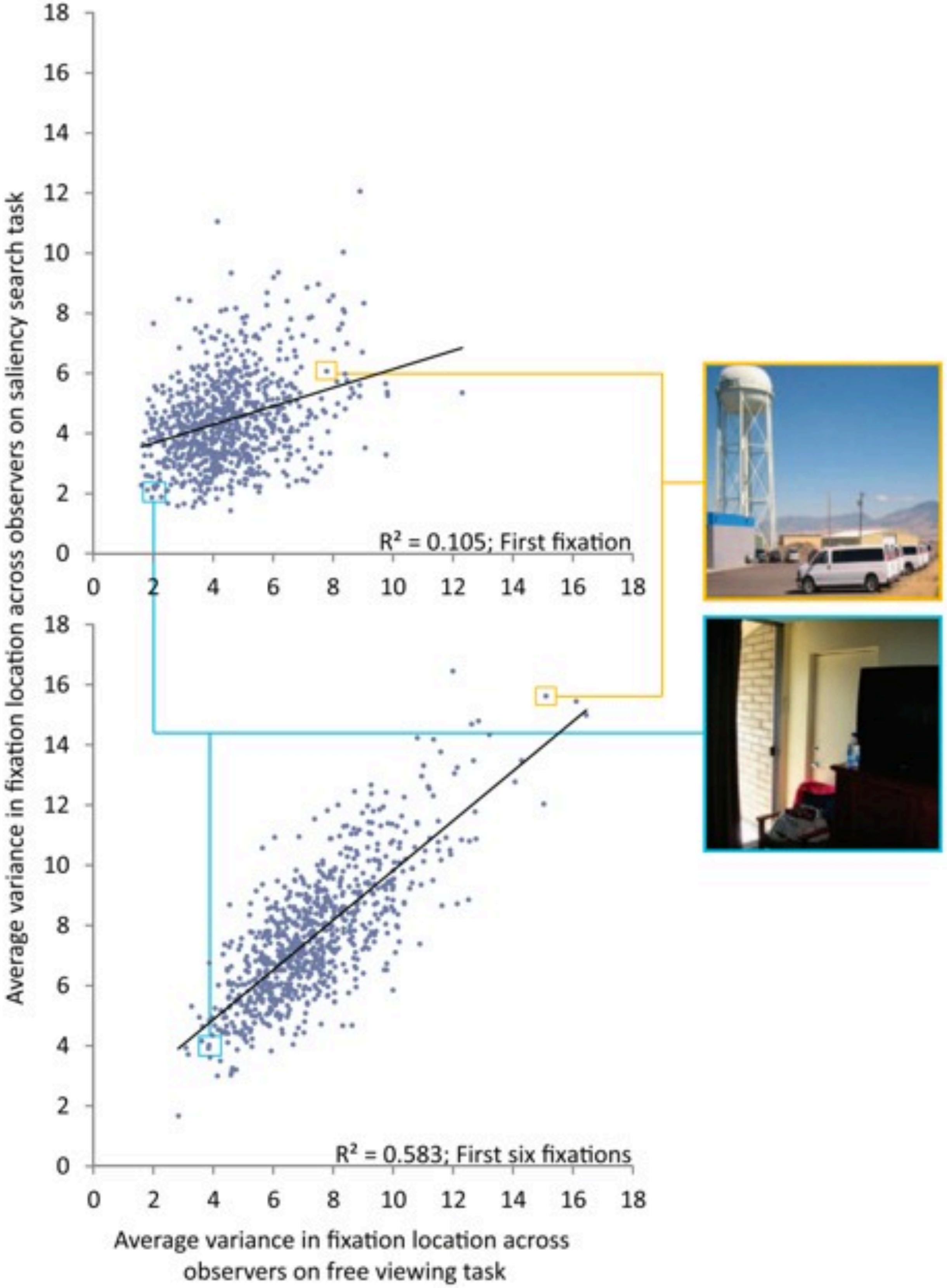
KL divergence and variability in location:
smaller for fixations than explicit judgements for the first 2 fixations,
comparable between fixations and explicit judgements for later fixations

F-test analyses (testing for differences in fixation variances) show: difference in variability between free viewing and object search, as well as between object search and saliency search; difference decreases as fixation number increases

similar trends with the F-test analyses for KL divergence



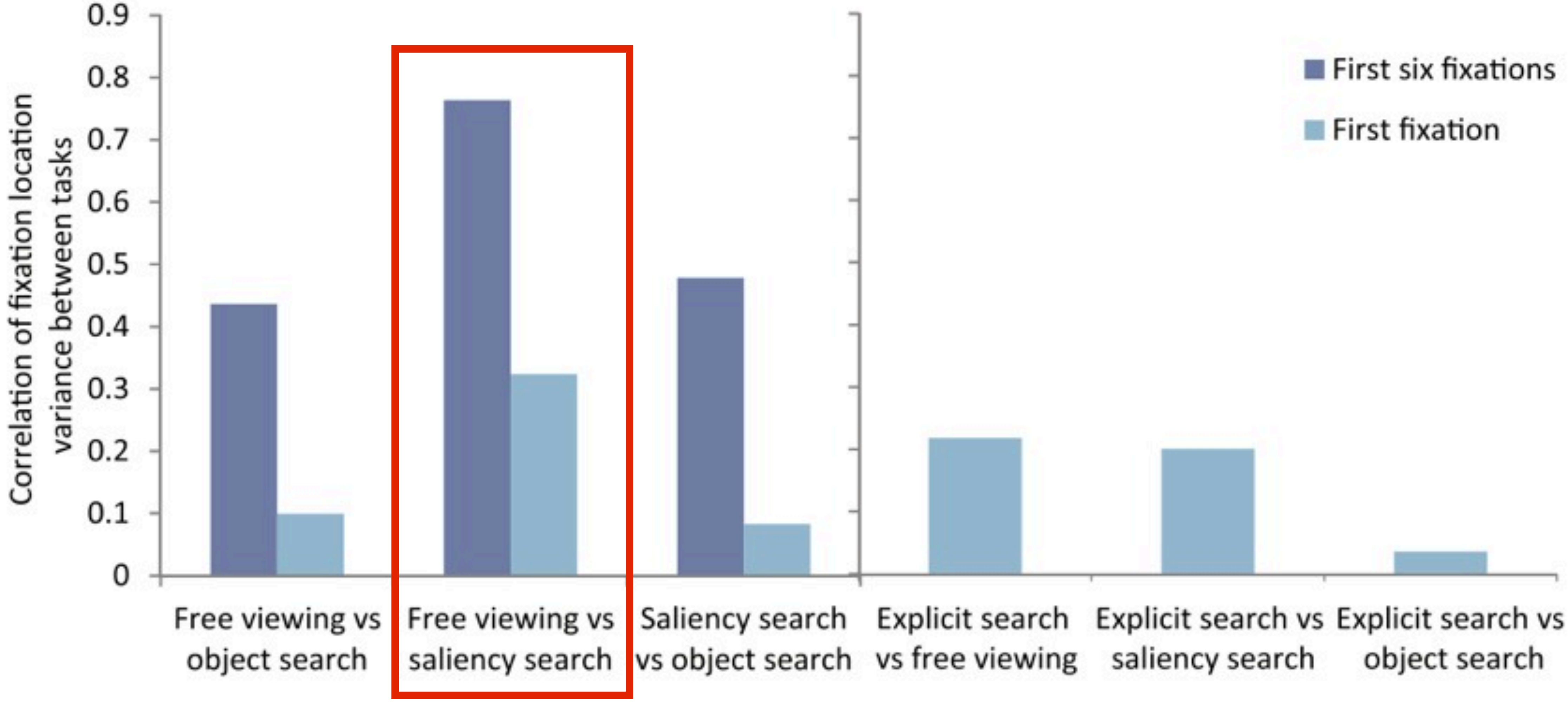
Correlation of observer variance across tasks



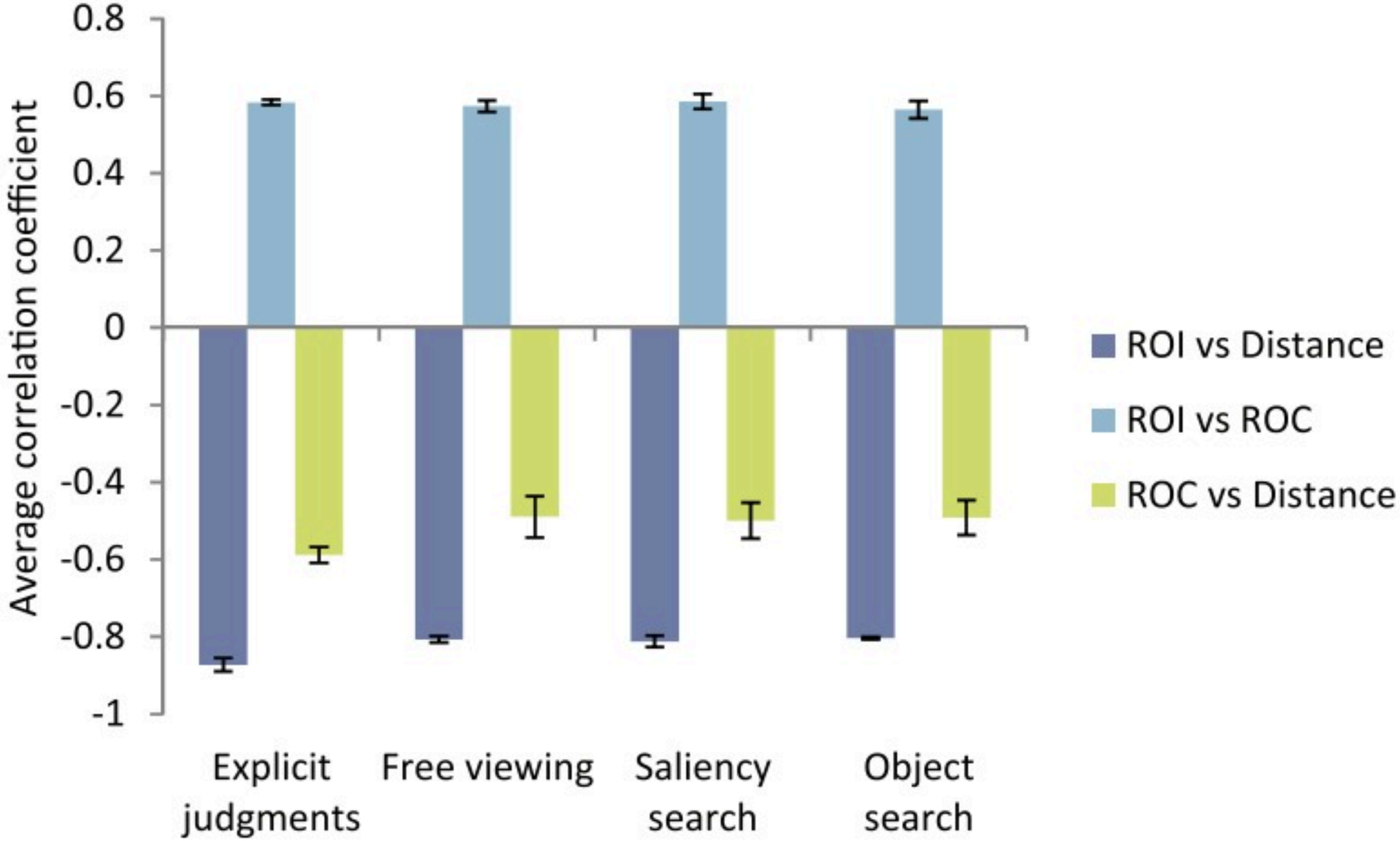
images containing many objects tend to produce highly variable eye movements

images with distinct salient objects tend to produce minimally variable eye movements

Correlation of observer variance across tasks shows: strongest correlation between free viewing and saliency search



Correlation of metrics across tasks shows:
ROI and distance metrics are highly correlated for all tasks
while both have lower correlation with ROC metric

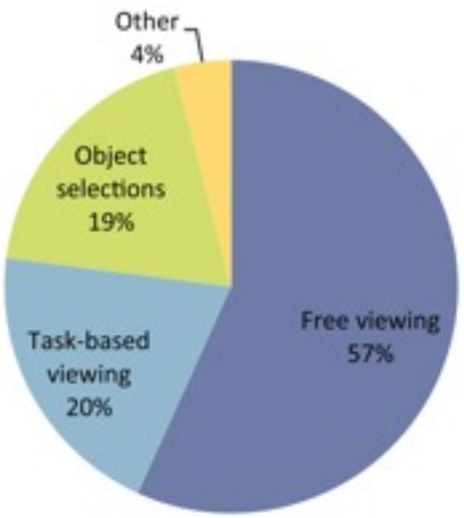


Conclusions about models + tasks

- all 3 saliency models tested were more accurate at predicting human judgments of explicit saliency than eye movements in the free viewing task
    - also, better model performance for saliency search than object search
- thus, free-viewing can not necessarily be described by directing eyes to bottom-up salient regions
    - different from explicit saliency judgments, so perhaps some other task in mind? (fixating important elements: faces, animals, text)
- but free-viewing is similar to moving eyes to salient locations
    - correlated predictions between free-viewing and saliency search tasks
    - perhaps share common strategy of fixating objects?
- so, saliency models can be used for modeling free eye movements, but with subpar performance
- results differ slightly depending on metric, because depends on how compatible models' output is with metric
    - e.g. maps that are too sparse do worse on ROC metric

Conclusions about observer variability

- eye movements across observers are more similar when they are instructed to engage in a specific task (object search vs free viewing)
- eye movements across observers during saliency search just as variable as for free viewing
  - more evidence that these tasks are performed similarly?
  - perhaps variability in what is considered a salient image region?
  - larger variation in later fixations (as opposed to earlier)
- inter-observer variability increases with fixation number
  - secondary idiosyncratic tasks engaged by each observer after task completion?
- cluttered images produce more differences in fixation locations across observers, and poorer predictions across models
  - want models that can be tested on complex images, and that can account for the full range of human behavior



Not the end of the story!

- many other computational saliency models
  - 35 models presented in Borji et al., 2012
- many other datasets
- many other tasks
  - see appendix of Koehler et al., 2014
- many other metrics
  - study of comparison metrics presented in Riche et al., 2014