Appendix A
Additional results

MIT1003 and CAT2000 results

Evaluations on the MIT1003 (Tab. A1) validations set corroborate the utility of pretraining with data generated from the EMMA cognitive model. Our approach improves over vanilla MD-SEM in all five metrics. Compared to recent SOTA approaches that report validation set results on MIT1003, our approach allows MD-SEM to reach competitive performance despite the significantly smaller number of parameters.

Our evaluation on CAT2000 (Tab. A2) is restricted to a comparison with MD-SEM, as reporting validation set results on CAT2000 is not common in previous work. However, the evaluation on CAT2000 is another clear indication of the benefit of pre-training with data generated by EMMA: our approach is superior to MD-SEM on all 5 metrics.

Category-specific results on CAT2000 validation set

Since the CAT2000 dataset is separated into 20 different image categories, we could evaluate the performance improvements of our approach compared to vanilla MD-SEM separately for each category (Tab. A3). Our approach is outperforms MD-SEM on every metric among all 20 categories, underlining the consistency of improvement using the cognitive model. Figure A1 and Figure A2 showcase some examples of the various categories. Such qualitative results illustrate that our approach has a more accurate estimation of salient regions than vanilla MD-SEM, especially in *Fractal*, *Jumbled*, *Low Resolution*, *Noisy*, and *Pattern*.

SALICON validation visualizations

In the main paper, Table 2 shows a comparison between our model's and current SOTA models' results on the SALICON test set. In order to qualitatively showcase how our approach performs compared to the vanilla MD-SEM, we pointed to 3. Here, in Figure A3 we show additional examples to give a better intuition of how pre-training with cognitive data helps to improve saliency prediction on SALICON.

Cognitive Pre-training vs. Computational Cost

Table 1 shows the improvement obtained by our training approach on the SALICON validation set on lightweight models, MSI-NET and MD-SEM. For completeness, we additionally provided the relative improvements gains between low versus high capacity models, showcasing that high capacity models do not benefit from cognitive model pre-training (Fig. A4). In addition, we provide a comparison in terms of absolute results of our experiments on the SALICON validation set in Tab. A4.

As discussed in the main paper, pre-training on EMMA-generated data leads to consistent improvements for models with a relatively low number of parameters (MSI-NET, MD-SEM), in contrast to the large EML-NET variants. Table A4 shows that EML-NET with DENSE and NASNET backbones still achieve the best NSS and SIM scores. However, such large model is outperformed by MD-SEM in AUC, CC and KL and in AUC and KL by MSI-NET when these models are pre-trained with EMMA-generated data.

Generalization on Unseen Data

To shed further light on the utility of using EMMA-generated data alone, we evaluated a variant of MD-SEM that is only trained on synthetic data and evaluated on ground truth MASSVIS (Tab. 5). In Table 5 we observe that when only training on FigureQA-EMMA without finetune on MASSVIS, the performance is worse than training on FigureQA and finetune on MASSVIS. The result indicates that training only on synthetic data is not sufficient. This is in line with previous work by Sood et al. (2020), who observed the same outcome when using synthetic reading behaviors from the EZ reader cognitive model to pre-train a text saliency prediction model. In their work, pre-training on EZ-generated data only led to worse results than the pure data driven approach. However, when pre-training with synthetic data and finetuning on ground truth human data, their saliency model produced best results. Figure C2 shows qualitative results on the MASSVIS test set.

Appendix B
Method

Extracting Synthetic Saliency Maps

We use the ACT-R Python implementation of EMMA.[10] The input for EMMA consists of a list of bounding box coordinates and the corresponding object labels. The object detection needs to cover the entire image and contents, since EMMA selects the next target from nearby bounding boxes. EMMA then outputs a scanpath as a list of coordinates and timings. In addition to those in Fig. 2, we show other examples of the generated saliency maps in Fig. B1. The code for generating saliency maps will be provided in our final release.

Estimating EMMA Parameters

Although cognitive models hold great promise to improve neural saliency prediction models, they were so far predominately applied to low complexity stimuli (Gobet, 1996). To apply the EMMA cognitive model to real-world images as well as information visualizations containing varied and potentially overlapping objects in complex scenes, we specifically optimized the models' parameters for each domain. In detail, we optimized the following EMMA parameters: eye movement scaling factor, eye movement angle, viewing distance, decay, retrieval latency and retrieval threshold. As the code for EMMA does not provide a built-in functionality for parameter optimization, we resorted to a grid search over the parameter space. We measured the quality of each set of parameters by computing the Earth Movers Distance (EMD) between the EMMA-generated and the ground truth saliency maps. To find the optimal EMMA parameters for natural images, we make use of the SALICON dataset. For information visualizations, we fit EMMA parameters on MASSVIS. The best parameters are shown in Tab. B2. Notice that values are different for

[10] https://github.com/jakdot/pyactr

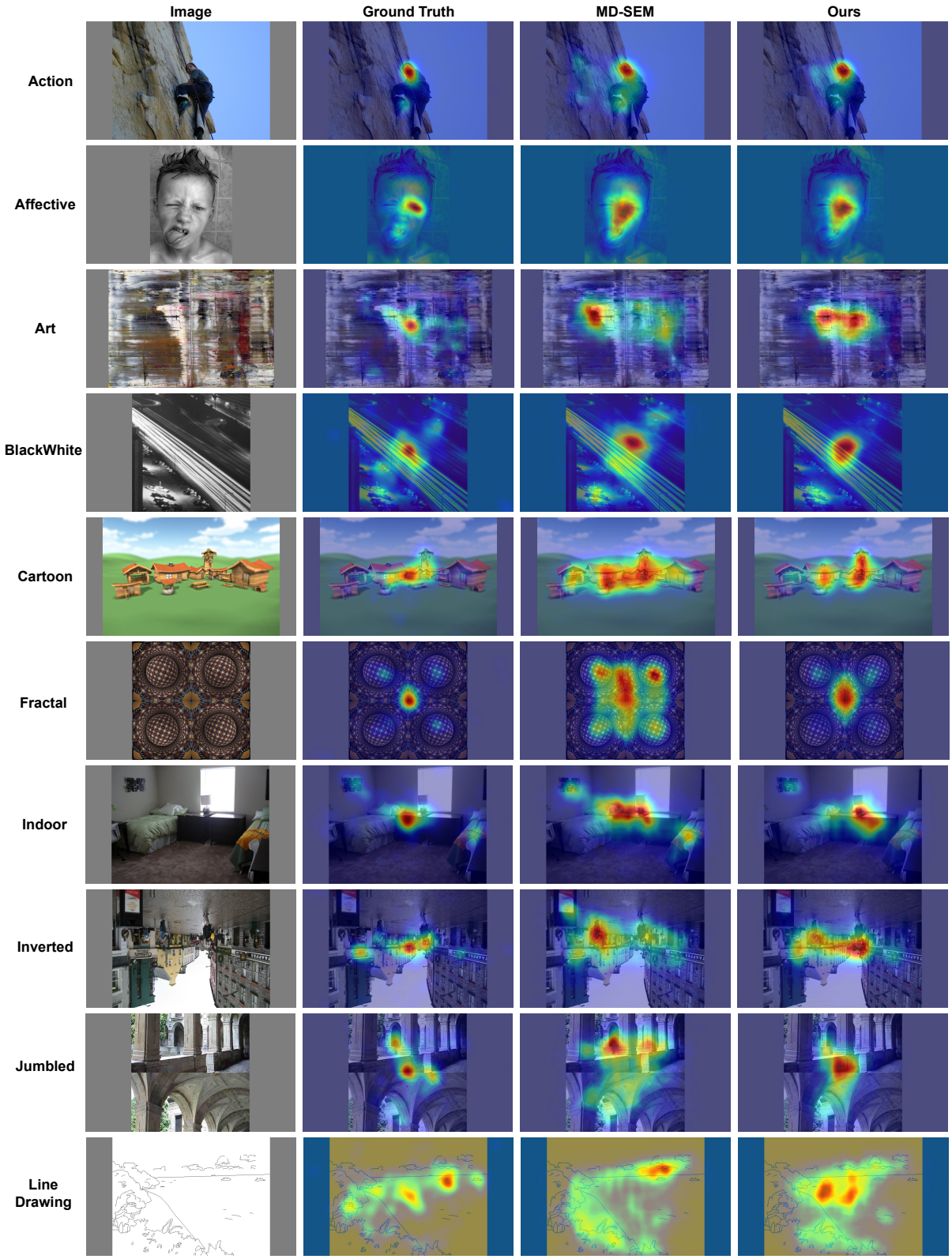

Figure A1. A visualization of example images from ten categories of CAT2000 validation set with the corresponding empirical ground truth maps, and predictions from our approach and the MD-SEM. The qualitative results indicate that the cognitive model enables the data-driven model to have a more accurate estimation of salient regions.

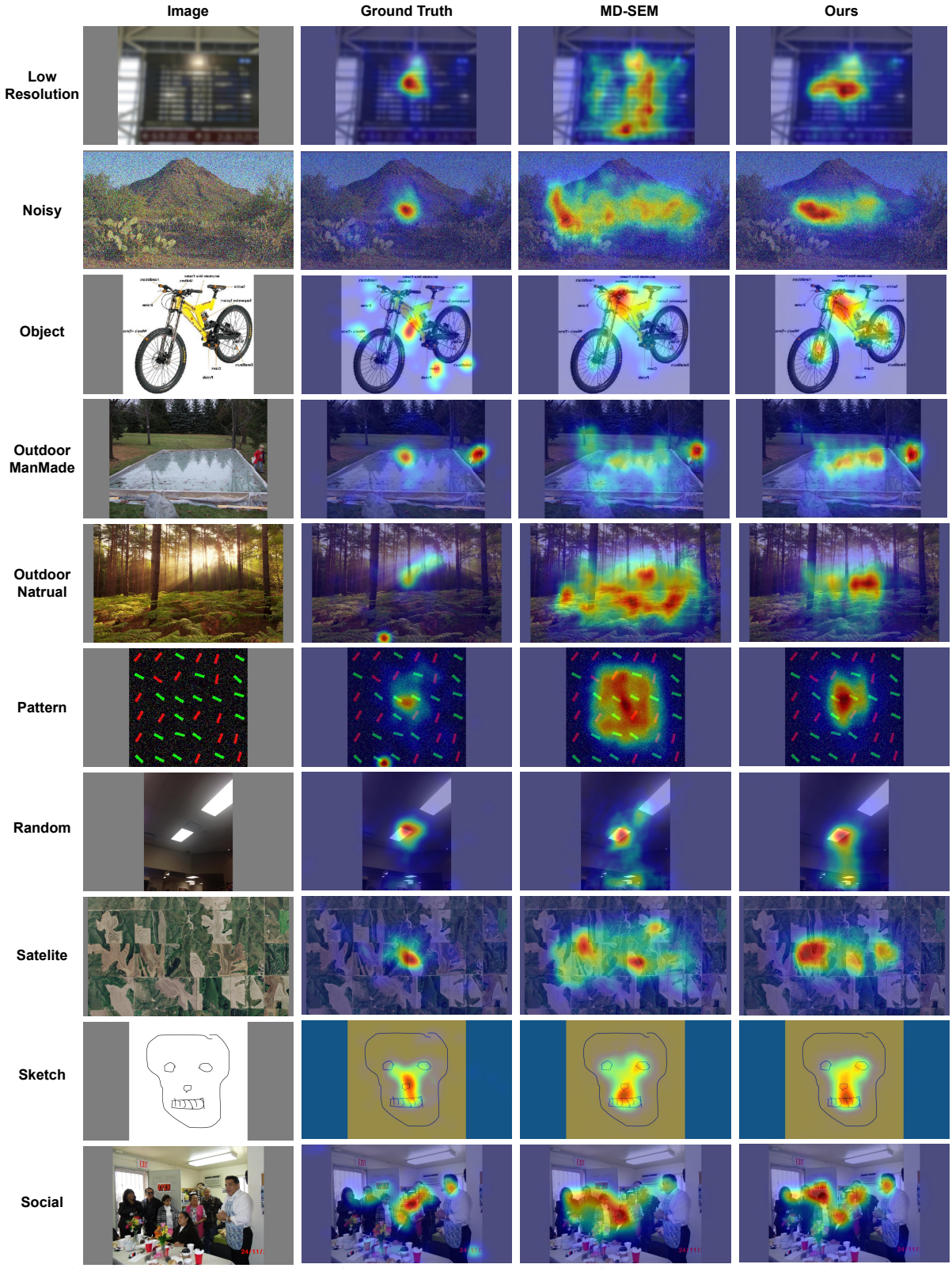

Figure A2. A visualization of example images from the other ten categories of CAT2000 validation set with the corresponding empirical ground truth maps, and predictions from our approach and the MD-SEM.

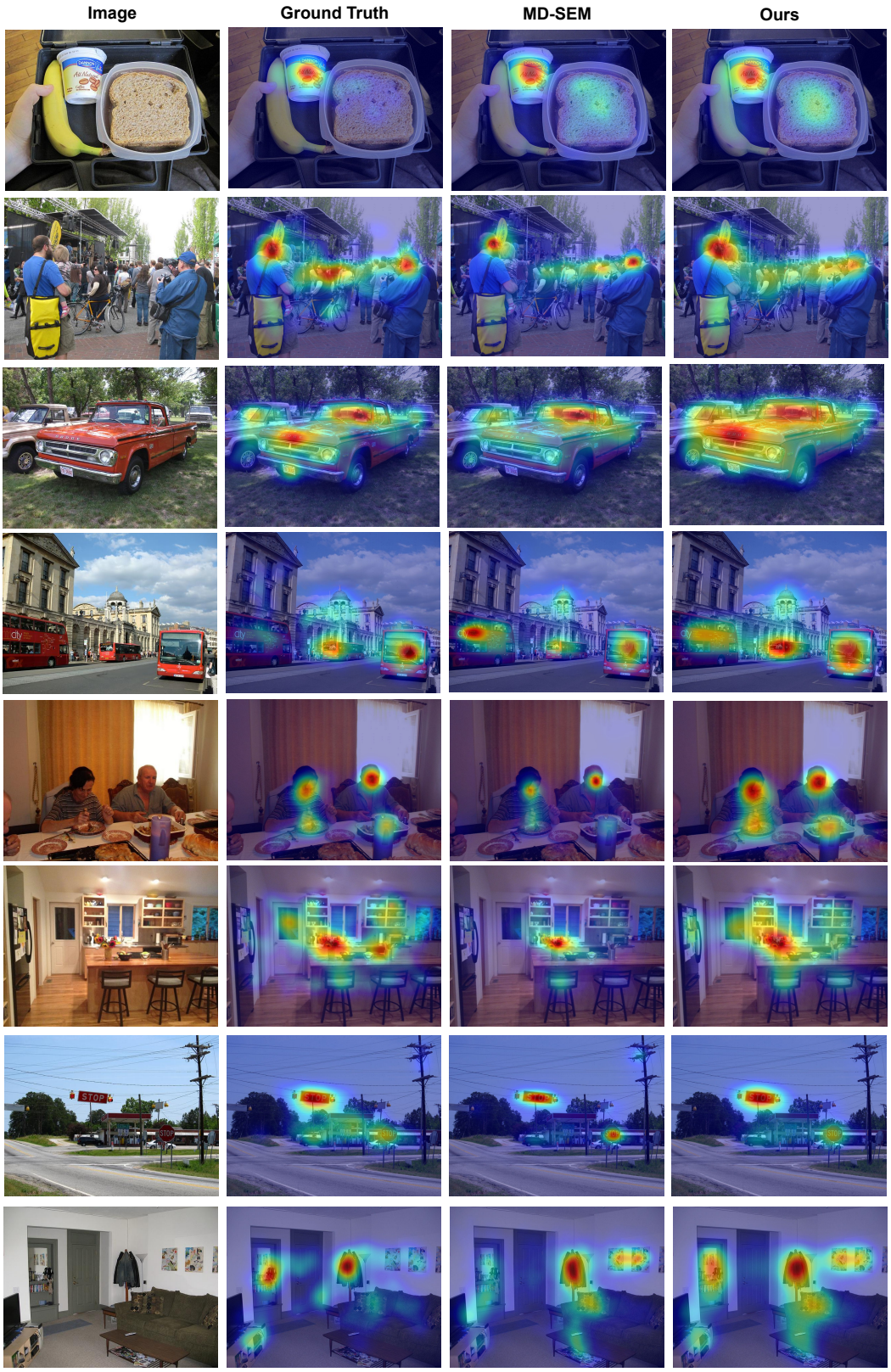

Figure A3. More example images from the SALICON validation set with the corresponding empirical ground truth maps, and predictions from our approach and the MD-SEM.

Table A1

*Prediction performance on MIT1003. The backbone network of DeepGazeIIE is densenet201. Best results are highlighted in **bold**, second best are underlined. Stars indicate statistical significance of the difference between Ours and MD-SEM (**: $p < .01$; ***: $p < .001$).*

Method	AUC ↑	CC ↑	KL ↓	NSS ↑	SIM ↑	Params
MSI-NET (Kroner et al., 2020)	0.832	0.741	0.818	2.663	0.602	24.9M
TranSalNET (Lou et al., 2022)	0.912	**0.774**	0.786	**2.921**	**0.628**	72.5M
SAM-ResNet (Cornia et al., 2018)	**0.913**	0.768	—-	2.893	—-	70.1M
DeepGaze IIE (Linardos et al., 2021)	0.889	**0.774**	**0.516**	2.599	—-	104.5M
MD-SEM (Fosco et al., 2020)	0.880	0.683	0.699	2.602	0.549	30.9M
Ours	0.897**	0.761**	0.544***	2.911**	0.609***	30.9M

Table A2

*Prediction performances of MD-SEM and our approach on the validation set of CAT2000. Best results are highlighted in **bold**. Stars indicate statistical significance of the comparison between Ours and MD-SEM (**: $p < .01$; ***: $p < .001$).*

Method	AUC ↑	CC ↑	KL ↓	NSS ↑	SIM ↑
MD-SEM (Fosco et al., 2020)	0.853	0.692	0.328	1.817	0.642
Ours	**0.867****	**0.782*****	**0.219*****	**2.074****	**0.694*****

the two datasets. Therefore, choosing the optimal parameters separately for each domain is crucial. In fact, if the best parameter set found on MASSVIS is applied to SALICON, that would result in an EMD of 1.96 instead of 1.48.

Our work showcases that by modifying parameters of EMMA, we can in fact improve the simulations across natural images as well as data visualizations to obtain more human-like attention according to the average EMD. These findings are not only interesting for the machine learning community as a first proof of concept that is possible to simulate human-like predictions on complex natural images and abstract visualizations, but it also indicates to cognitive science researchers that there is potential to improve EMMA predictions on such stimulus domains as we for the first time evaluate EMMA prediction performance on complex, real-world stimuli.

EMMA vs. Random Gaze

Table B1 shows a comparison between EMMA-generated and randomly generated gaze data on SALICON and MASSVIS. As reported in the main paper, EMMA-generated gaze data results in better CC, KL and SIM scores than the randomly generated gaze.

Appendix C
Synthetic gaze datasets

We release two large scale synthetic gaze datasets, MSCOCO-EMMA and FigureQA-EMMA. MSCOCO-EMMA contains EMMA-generated scanpaths and saliency maps on the MSCOCO dataset, based on the optimal EMMA parameters estimated on SALICON. FigureQA-EMMA contains EMMA-generated scanpaths and generated saliency maps on the FigureQA dataset based on EMMA parameters estimated on MASSVIS. Our novel synthetic datasets are significantly larger compared to current human saliency datasets (130k images in MSCOCO-EMMA versus e.g. 15k images in SALICON) and can be a valuable resource for training future neural saliency prediction models. Table C1 summarizes the details of our synthetic datasets.

In Figure C1, we show visualizations of samples from MSCOCO-EMMA and FigureQA-EMMA. According to different parameter sets, the number of bounding boxes, and sizes of objects in the two datasets (see Figure B2 and Table B2), EMMA outputs human-like attention data which is specific to each of the domains. In Table B2 we show the best set of parameters differs between dataset domains.

We further analyze the impact that the number of objects in each respective domain has on said parameters (e.g., viewing distance, eye movement angle, etc). In Figure B2, we visualize the relationship between the number of objects per image and performance of EMMA in terms of Earth Movers' Distance (EMD). The pattern we observe differs between natural images (SALICON) and information visualizations (MASSVIS) domains. As the number of objects increases, the mean EMD decreases for natural images (SALICON). However, when there are more than 60 objects in the scene, the mean EMD increases. For data visualizations (MASSIVS), for more than 20 objects, as the number of objects increases the mean EMD increases. We performed separate linear regressions, which were significant on both datasets with $p < 0.001$. For SALICON, the linear regression indicated an overall positive relationship between number of bounding boxes and EMD, whereas for MASSVIS this relationship is reversed.

Appendix D

Table A3

Performance improvements of our approach over vanilla MD-SEM for 20 categories from the CAT2000 validation set. Stars indicate statistical significance of the comparison between Ours and MD-SEM (: $p < .05$, **: $p < .01$; ***: $p < .001$).*

	AUC ↑	**CC ↑**	**KL ↓**	**NSS ↑**	**SIM ↑**
Action	+0.004	+0.031	-0.035	+0.126	+0.026
Affective	+0.010	+0.045	-0.098	+0.169	+0.053
Art	+0.008	+0.062	-0.055	+0.138	+0.033
BlackWhite	+0.008	+0.062	-0.055	+0.138	+0.033
Cartoon	+0.021	+0.104*	-0.108*	+0.237	+0.050*
Fractal	+0.002	+0.092	-0.104	+0.302	+0.049
Indoor	+0.021	+0.103*	-0.118***	+0.236	+0.048**
Inverted	+0.022	+0.080	-0.105*	+0.177	+0.050*
Jumbled	+0.025	+0.145**	-0.127*	+0.332*	+0.048**
LineDrawing	+0.018	+0.097*	-0.119**	+0.263	+0.059
LowResolution	+0.021	+0.192***	-0.260***	+0.643**	+0.120***
Noisy	+0.010	+0.091	-0.106	+0.267	+0.047
Object	+0.005	+0.041	-0.056	+0.182	+0.033
OutdoorManMade	+0.013	+0.072	-0.081	+0.184	+0.038
Pattern	+0.007	+0.098**	-0.102*	+0.265	+0.058*
Random	+0.015	+0.011	-0.138	+0.323	+0.060
Satelite	+0.014	+0.072	-0.078	+0.177	+0.029
Sketch	+0.005	+0.026	-0.037	+0.085	+0.030
Social	+0.016	+0.078	-0.093	+0.242	+0.056**

Table A4

*Saliency prediction performance several SOTA models on the SALICON validation set, with and without pre-training on synthetic data from EMMA. As well as pre-training only, and evaluated on groundtruth data – without finetuning (FT) on target dataset. Best results are highlighted in **bold**, second best underlined, and third best in italic.*

Methods	**AUC ↑**	**CC ↑**	**KL ↓**	**NSS ↑**	**SIM ↑**
MSI-NET	0.855	0.840	0.265	1.784	0.740
MSI-NET with EMMA	**0.866**	0.886	0.198	1.891	*0.776*
EML-NET (DENSENET)	0.797	0.873	0.230	1.987	0.769
EML-NET (DENSENET) with EMMA	0.858	0.887	0.250	*1.972*	0.775
EML-NET (DENSE + NASNET)	0.802	*0.890*	*0.204*	**2.024**	**0.785**
EML-NET (DENSE + NASNET) with EMMA	*0.861*	0.891	0.232	1.914	0.780
MD-SEM	0.858	0.843	0.268	1.818	0.732
MD-SEM with EMMA (no FT)	0.660	0.680	0.576	1.455	0.580
MD-SEM with EMMA (ours)	0.865	**0.894**	**0.193**	1.891	0.780

Table B1

Comparison of EMMA-generated gaze with randomly generated gaze on SALICON and MASSVIS. Best results are shown in bold.

Dataset	**Method**	**CC ↑**	**KL ↓**	**SIM ↑**
SALICON	EMMA	**0.432**	**1.624**	**0.457**
	Random Gaze	0.167	2.068	0.332
MASSVIS	EMMA	**0.4**	**0.641**	**0.586**
	Random Gaze	0.096	1.703	0.307

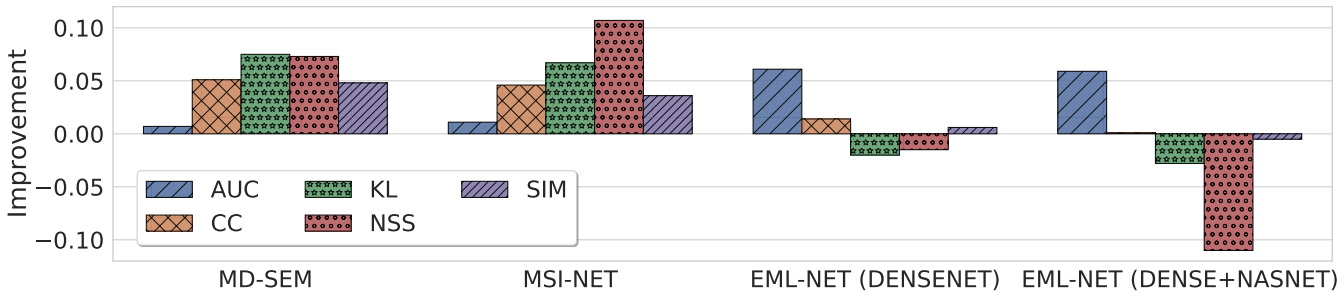

Figure A4. Prediction performance of our approach versus data-driven only models on the SALICON validation set. A bar above zero means an improvement in the respective metric.

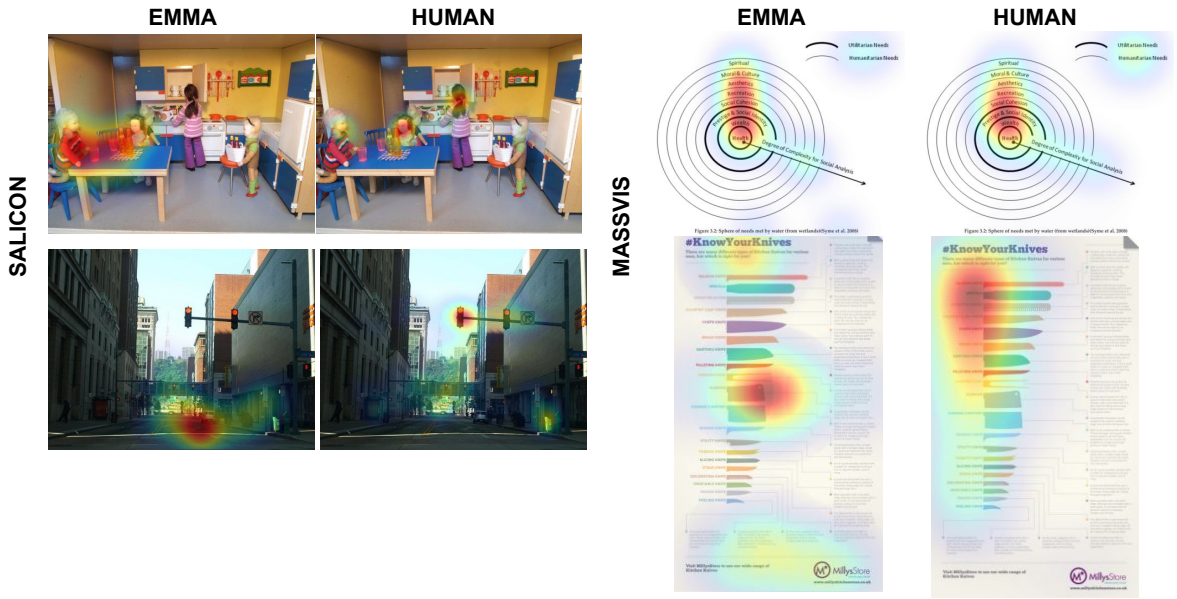

Figure B1. More example images from the SALICON and MASSVIS of synthetic EMMA-based gaze maps versus human ground truth maps.

Training details

Training details

For comparability, we made use of the official MSI-NET and MD-SEM code provided by the authors. [11] The only change introduced by our approach is that we pre-trained on synthetic gaze data from the EMMA cognitive model, and then finetuned with standard ground truth human attention data. To analyze the generalizability of our approach, we performed saliency prediction on two domains: Natural images and information visualizations. All experiments were conducted on a single NVIDIA Tesla V100 GPU with 32 GB VRAM.

For all four datasets, we set the batch size to 8 and the initial learning rate to 10^{-5}, reducing it by a factor of ten every two epochs. Models are trained on MSCOCO-EMMA for 5 epochs, on SALICON for 13 epochs, on CAT2000 for 9 epochs, and on MIT1003 for 15 epochs. We split MSCOCO-EMMA into 70% train and 30% for validation. For SALICON, we used the provided splits as 10,000 images in train and 5,000 images in validation set. We split the CAT2000 and MIT1003 data according to previous work (Jia & Bruce, 2020; Kroner et al., 2020). For CAT2000, this results in 1,800 images (randomly selected 90 per category) for training and 200 for validation (rest 10 per category). For MIT1003, in accordance to (Cornia et al., 2018), we obtain 903 training and 100 for test samples.

We used a batch size of 8 and an initial learning rate of 10^{-5}. We pretrained on FigureQA-EMMA for 15 epochs, and then fine-tuned on MASSVIS for 8 epochs. For FigureQA, we split the data into 70% for training and 30% for validation. For MASSVIS, we split the data into 70%, 10% and 20% for train, validation and test respectively.

Appendix E
Limitations and future work.

While the EMMA cognitive model led to consistent performance improvements, other cognitive models might be even

[11] https://github.com/diviz-mit/multiduration-saliency, https://github.com/alexanderkroner/saliency

Table B2
Best parameter set for EMMA on natural images and on data visualization images and average number of bounding boxes in SALICON and MASSVIS.

Parameters	SALICON	MASSVIS
Eye movement scaling factor	0.1	0.001
Eye movement angle	25	20
Viewing distance	120	140
Decay	0.5	0.5
Retrieval latency	0.4	0.4
Retrieval threshold	-2	-2
Images	15000	393
Avg.BBox	37	22
EMD ↓	1.4835	1.4166

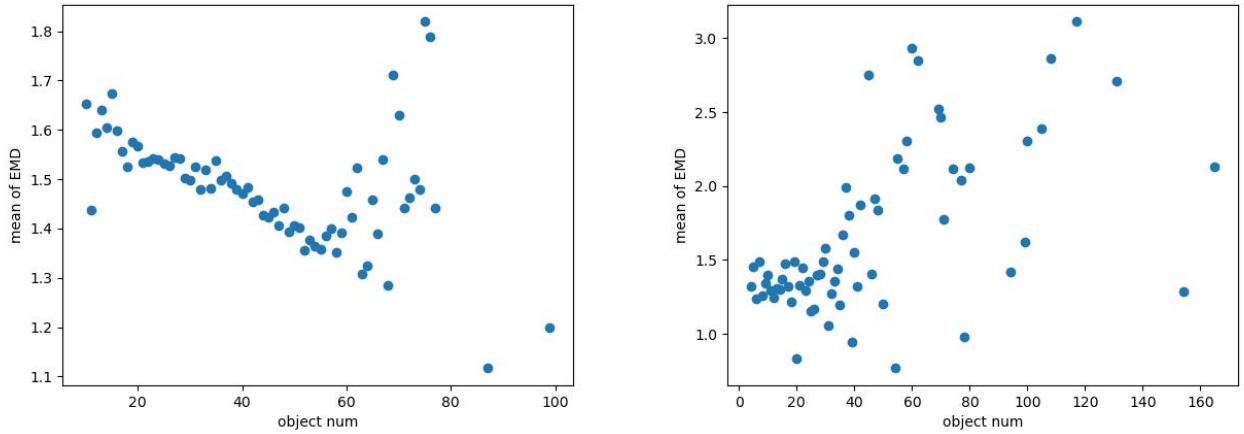

Figure B2. The number of objects versus mean EMD score. Left: SALICON. Right: MASSVIS

Table C1
Statistics for MSCOCO-EMMA and FigureQA-EMMA.

	#Images	#Scanpaths	#Saliency Maps	#Avg.BBox
MSCOCO-EMMA	130k	130k	130k	32
FigureQA-EMMA	100k	100k	100k	79

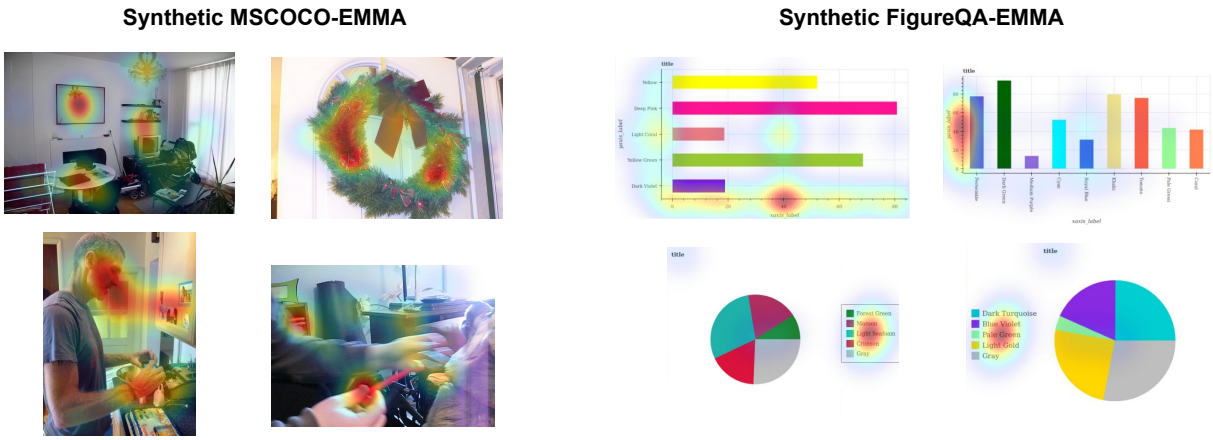

Figure C1. A visualization of example images from our synthetic datasets, MSCOCO-EMMA and FigureQA-EMMA.

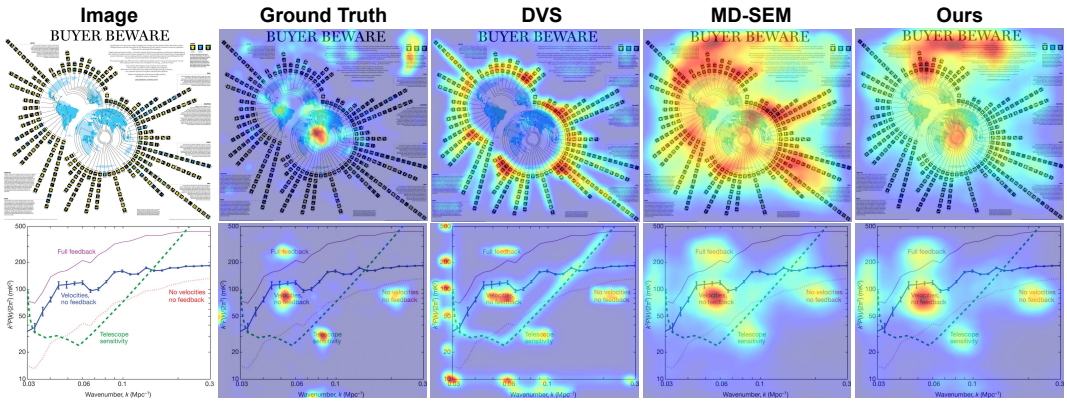

Figure C2. A visualization of example images from the MASSVIS dataset with the corresponding empirical ground truth maps, predictions from our approach and baseline methods.

better suited for the task. In future work, we plan to address this limitation by comparing the utility of different cognitive models (Kieras & Meyer, 1994; Nyamsuren & Taatgen, 2013b; Rohrer, 2008) when integrated in the training process of neural saliency prediction approaches. Another limitation to our work is the use of SALICON to find the optimal parameters of EMMA. SALICON provides ground truth saliency maps obtained from mouse tracking data. These saliency maps were proven to be highly effective for pretaining of saliency prediction networks and also resulted in highly useful parameters for EMMA. However, it is possible that the use of real gaze data for the EMMA parameter search might lead to even stronger performance. Lastly, in our work we focused on free-viewing scenarios to prove the utility of training with a cognitive model of visual behavior. EMMA, however, is also capable of producing goal-directed behavior. Future work should explore the capability of EMMA under other paradigms, like visual search. It would also be interesting to extend our approach to dynamic stimuli to improve video saliency prediction (Droste et al., 2020) or even attention prediction in interactive environments (Vozniak et al., 2022).