Supplementary Material
How much time do you have? Modeling multi-duration saliency

Camilo Fosco[*1], Anelise Newman[*1], Pat Sukhum[2], Yun Bin Zhang[2], Nanxuan Zhao[3], Aude Oliva[1], and Zoya Bylinskii[4]

[1]Massachusetts Institute of Technology, [2]Harvard University, [3]City University of Hong Kong, [4]Adobe Research

{camilolu, apnewman, oliva}@mit.edu, {psukhum, ybzhang}@g.harvard.edu, nanxuanzhao@gmail.com, bylinski@adobe.com

The supplement contains additional details about the CodeCharts1K dataset, the architecture of MD-SEM, and the effectiveness of CCM loss, as well as additional results.

1. CodeCharts Validation Procedure

Our validation procedure for CodeCharts1K was designed to eliminate data from inattentive or noncompliant participants. We follow the validation procedure in [2]. We include validation images consisting of a cropped human face on a plain background, and participants are expected to enter a code that overlaps the validation cue (the face image). We discard data from participants who miss more than 25% of validation images. We also discard data from participants who look at the same spot on an image (within a radius of 100 pixels) for at least five images in a row, to eliminate people who consistently fixate at the same spot. For CodeCharts1K, we collected 50 gaze points per image per duration; after data filtering, we had on average 44 gaze points (we discarded approximately 12% of the data). These validation procedures were implemented for the CodeCharts1K dataset based on observations during our pilot study on the OSIE data, which is why we see lower inter-observer consistency (IOC) on OSIE.

2. Data analysis

2.1. Do people look in the same places?

To judge whether people look at the same image regions at different durations, we ran a split-half consistency analysis using different dataset subsets from CodeCharts1K. For each image and viewing duration, we generate 10 splits of participant data by resampling gaze points for a given duration. We use these gaze points to generate a heatmap and use it to compute the Pearson's Correlation Coefficient (CC) score between heatmaps at different durations. The final numbers in Fig. 1 are produced by averaging over im-

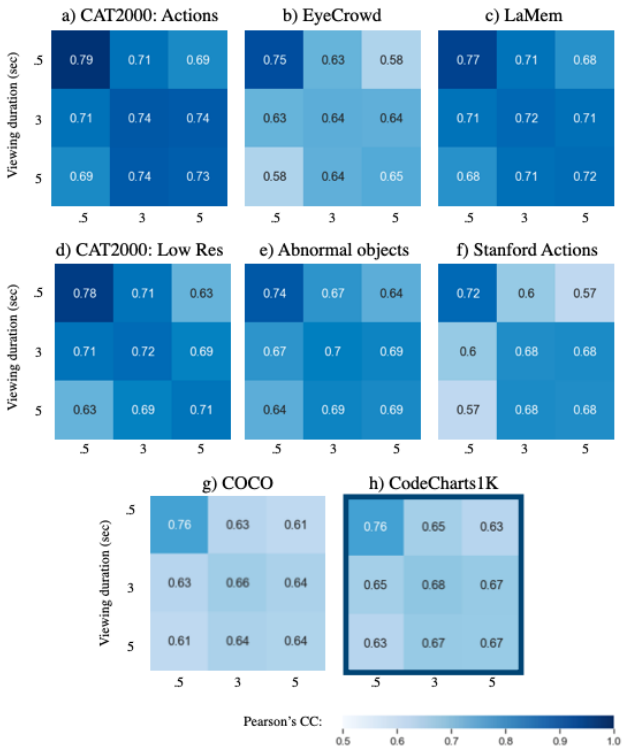

Figure 1. Split-half consistency of viewers, computed using Pearson's Correlation Coefficient (CC), within and across viewing durations on different subsets of CodeCharts1K (a-g) and on the full dataset (h).

[*]Equal contribution.

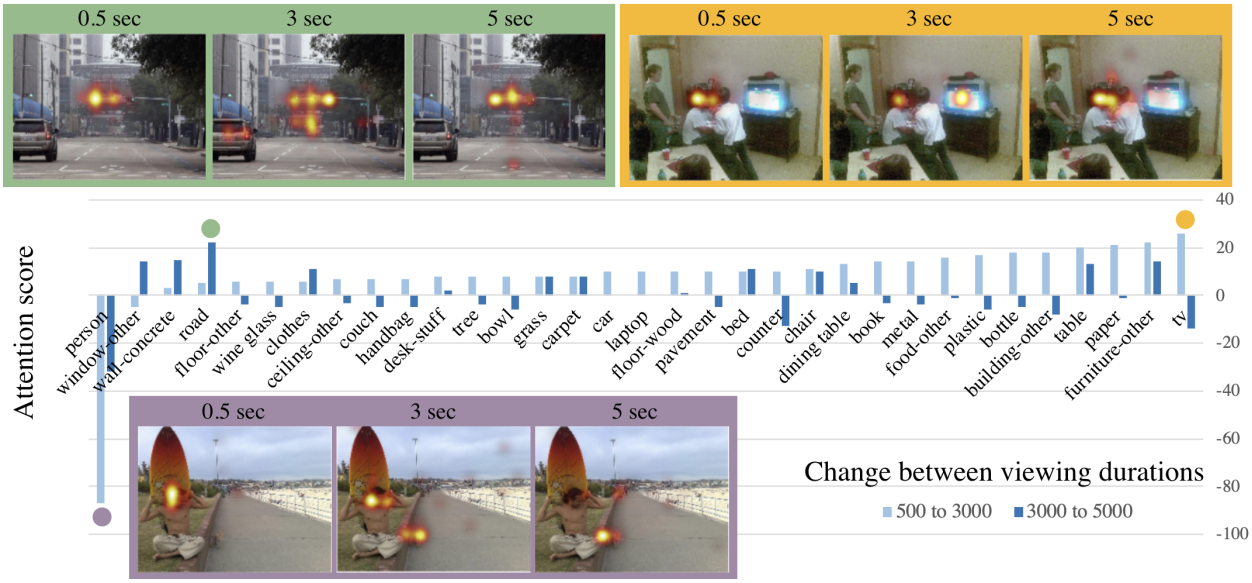

Figure 2. People's gaze falls on different things and "stuff" in images over time. From 0.5 sec to 3 sec, gaze frequently moves away from people and is pulled towards objects and furniture. From 3 to 5 sec, gaze on contextual "stuff" like grass, carpet, pavement, and road increases. COCO segmentations [1] were used for this analysis. An object in an image was assigned a score of +1, -1, or 0 depending on whether it increased, decreased, or didn't change in saliency from one duration to the next. These scores were added up over all the images where the objects occurred to produce the attention scores plotted on the y-axis. Example images containing patterns of gaze change over time for the people, road, and tv object categories are included above.

ages and splits. These results confirm that consistency is high within participants viewing images at a particular duration (diagonal entries), and that consistency is highest at the shortest duration. Gaze patterns at 3 and 5 seconds are frequently quite similar to each other, however. We note that these split-half consistency analyses aggregate the gaze points of only 18-22 participants per group as opposed to the 44 gaze points on average in the full dataset; therefore, the consistency numbers in Fig 1 underestimate the robustness of the full data.

2.2. What is salient at what time?

We used COCO segmentation maps [1] available for the SALICON images in our dataset to compute gaze counts per image segment across time (Fig. 2). We compute an "attention score" per object by giving the object a score of +1 every time it increases in saliency from one viewing duration to the next, a score of -1 if it decreases, and 0 otherwise. We sum these scores across the images in our dataset. The benefit of this score is avoiding image-specific saliency scale differences. We find that from 0.5 sec to 3 sec, gaze frequently moves away from people and is pulled towards objects and furniture (e.g., TVs, tables, bottles, chairs, books, etc.). From 3 to 5 sec, there is an increase of attention on contextual "stuff" like roads, walls, windows that may contain other objects. At these longer durations people notice smaller and more distant objects in an image.

	Input	Output
GAP	-	2048
Dense	2048	512
LSTM	512	512
Dense	512	2048
Sigmoid	2048	2048

Table 1. The architecture of temporal excitation module.

3. Model architecture

Tables 1 and 2 lay out the architecture of the custom modules of MD-SEM. Table 1 covers the Temporal Excitation Module and Table 2 covers the decoder.

4. Effectiveness of CCM Loss

The Correlation Coefficient Match (CCM) loss explicitly encourages our network to model temporal differences in saliency data. Table 3 shows how adding CCM to our loss function improves the performance of MD-SEM and SAM-MD on CodeCharts1K.

5. Additional evaluations and predictions

Fig. 3 compares predictions from MD-SEM to other models (SAM-MD and SAMx3). Fig. 4 shows representative predictions of MD-SEM on various datasets.

	Kernel	Stride	Dialation	Output
Conv.	3 x 3	1 x 1	2 x 2	256
Conv.	3 x 3	1 x 1	2 x 2	256
Upsample	-	2 x 2	-	256
Dropout (0.3)	-	-	-	256
Conv.	3 x 3	1 x 1	2 x 2	128
Conv.	3 x 3	1 x 1	2 x 2	128
Upsample	-	2 x 2	-	128
Dropout (0.3)	-	-	-	128
Conv.	3 x 3	1 x 1	2 x 2	64
Conv,	3 x 3	1 x 1	2 x 2	64
Upsample	-	2 x 2	-	64
Dropout (0.3)	-	-	-	64
Conv.	1 x 1	1 x 1	1 x 1	3

Table 2. The architecture of the decoder.

Model	*NSS* ↑	*CC* ↑	*KL* ↓	*SIM* ↑	*CCM* ↓
SAM-MD w/o CCM	2.700	0.744	0.434	0.616	0.231
SAM-MD w/ CCM	2.739	0.753	0.458	0.609	0.198
MD-SEM w/o CCM	2.778	0.754	0.565	0.598	0.228
MD-SEM w/ CCM	**2.915**	**0.765**	**0.430**	**0.620**	**0.195**

Table 3. MD-SEM results on CodeCharts1K with and without CCM loss. We report performance on NSS, CC, KL, SIM and our custom CCM loss. These results correspond to the average over all durations.

In Table 4, we show MD-SEM performance on the different datasets that compose CodeCharts1k. Our model performs well in situations with humans and memorable objects, but struggles in images with uncommon scenes, complex actions or out-of-context objects, as the attention patterns are more affected by higher-level cognitive effects. These results highlight potential directions for future work.

Dataset	*NSS* ↑	*CC* ↑	*KL* ↓	*SIM* ↑
Stanford-Actions	2.698	0.710	0.493	0.594
EyeCrowd	2.728	0.765	0.434	0.611
Out-of-context + Abnormal	2.767	0.755	0.432	0.613
SALICON	2.908	0.758	0.447	0.613
CAT2000	3.090	0.791	0.373	0.646
LaMem	3.118	0.796	0.388	0.643

Table 4. MD-SEM results on the different sub-datasets that compose CodeCharts1K.

6. Applications

Fig. 5 contains examples of how multi-duration saliency heatmaps can be used to crop parts of an image that are salient at different times. Fig. 6 shows how multi-duration saliency can be used to select which elements in an image should be rendered first (or at higher resolution). Fig. 7 contains additional examples of how multi-duration saliency can be used to generate captions that pick up on additional objects or focus on salient parts of an image.

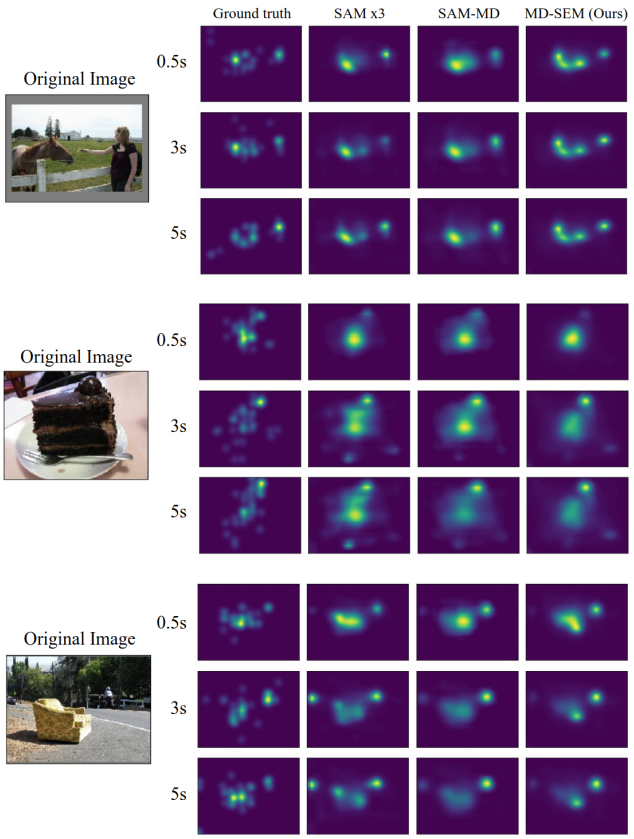

Figure 3. Comparison of saliency predictions from MD-SEM, SAM-MD and 3 SAMs individually trained to predict saliency at different durations. (a) Our model approximates shifts in attention more consistently on longer durations than SAMx3 and SAM-MD, here capturing the focus on the horse at longer durations in a more precise manner. (b) SAMx3 and SAM-MD struggle to generate precise heatmaps, while our model accurately predicts the focus in attention on the berry at 3 and 5 seconds. (c) While SAMx3 and SAM-MD correctly predict some punctual attention spots (top right corner at 3 seconds), they fail to recognize that gazes tend to return to the initial object of attention on longer durations.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Computer vision and pattern recognition (CVPR), 2018 IEEE conference on*. IEEE, 2018.

[2] Anelise Newman, Barry McNamara, Camilo Fosco, Yun Bin Zhang, Pat Sukhum, Matthew Tancik, Nam Wook Kim, and Zoya Bylinskii. TurkEyes: A web-based toolbox for crowdsourcing attention data. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

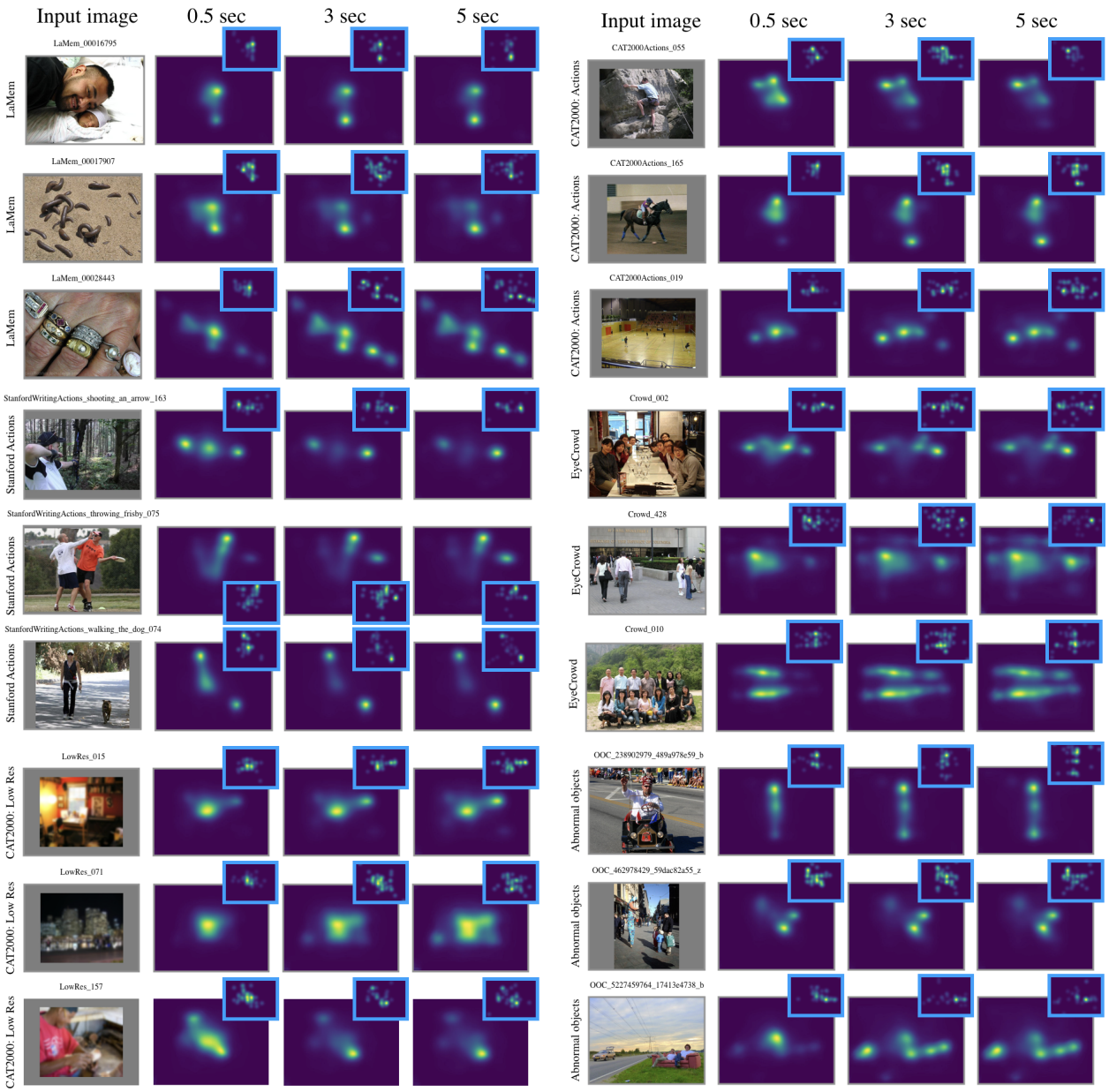

Figure 4. Saliency predictions of MD-SEM on various datasets. Insets with blue borders contain human ground-truth gaze locations collected using our CodeCharts UI. In all cases, we see that our model has learned to make distinctively different predictions for the different viewing durations. The model learns to start either more centrally or by focusing on the main actor in the scene in the first 0.5 sec. With longer viewing durations the model's predictions move towards other salient image elements that are smaller or more distant from the center. We can see a failure mode of our model on the large crowd of people in the right column, as our model struggles to determine who to focus on. We see another two difficult cases in the last two rows of the same column, where the model needs semantic knowledge to correctly distribute attention to objects that are out of place.

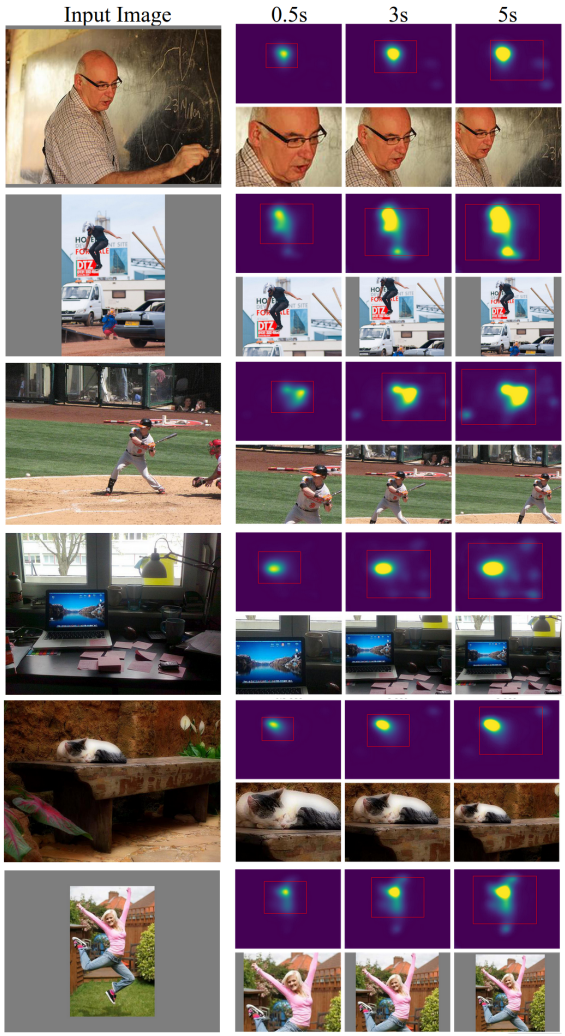

Figure 5. Example crops generated based on saliency maps for different viewing durations. The original images appear on the left. On the right we show the predicted saliency heatmap, along with the 90% bounding box, for each duration (top row) and the resulting cropped image (bottom row). Crops for 0.5 seconds tend to focus on a single highly salient object or point, while crops at longer durations expand to include other parts of the image such as the background or the object of the action.

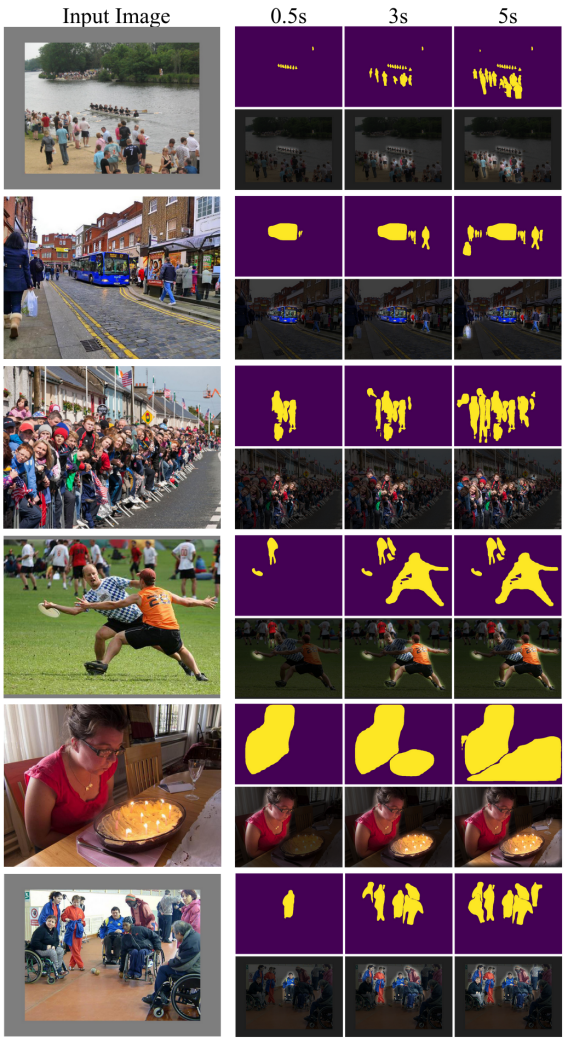

Figure 6. Examples of how multi-duration saliency can be applied to compression and rendering. The original images appear on the left. On the right we show the segmentation maps of instances with saliency scores in the 90th percentile based on the cumulative saliency map for that duration (top row) and a visualization of those salient objects (bottom row). Objects that are highly salient at 0.5 or 3 seconds could be rendered before objects that become salient later.

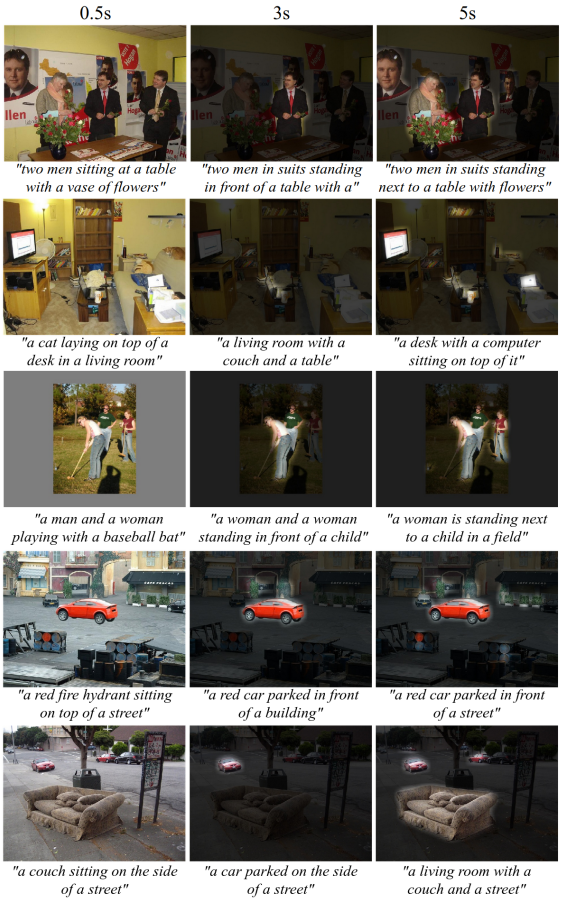

Figure 7. Examples of how multi-duration saliency can be applied to captioning. The captions corresponding to saliency-enhanced images for different durations can sometimes produce different captions by refocusing attention on relevant areas in a scene.