

Human vs. AI: A Novel Benchmark and a Comparative Study on the Detection of Generated Images and the Impact of Prompts

Philipp Moeßner, Heike Adel
Hochschule der Medien, Stuttgart, Germany



Motivation

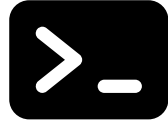


Possible
disinformation



Need for reliable
detection methods

Motivation



Impact of prompt
on detection
performance?

Sha et al. (2023)
show specific words /
prompt lengths lead
to lower detection
performance

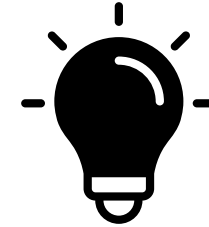
**No detailed
investigation, no
human performance**



Evaluation of
**human and machine
performance?**

User studies: Cooke et al.
(2024), Lu et al. (2024);
Machine evaluation: Park et
al. (2024), Corvi et al. (2024)

**No direct comparison
possible → different
test images**



Evaluation of
human and machine
decision rationales?

For humans:
Pocol et al. (2023);
For models: Bird
and Lotfi (2024)

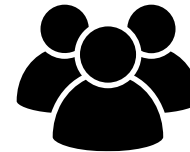
**No direct comparison
possible → different
evaluation methods**

Research Questions and Contributions

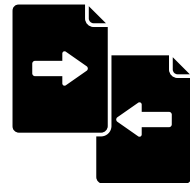
1. Does the prompt's level of detail have an impact on detection performance?
2. Does human or AI-model performance dominate?
3. Do they consider the same objects and structures in an image when evaluating it?



COCOXGEN
novel benchmark dataset

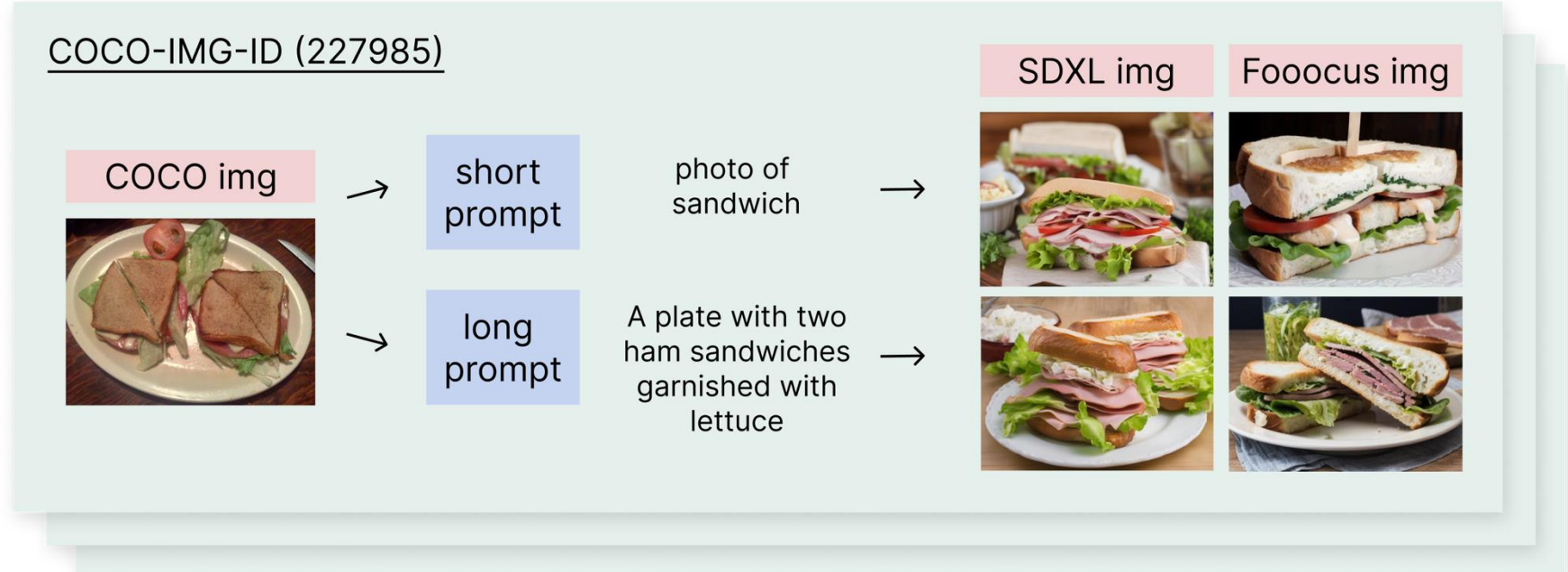


Conduction of a
large-scale user study



Direct **comparison of human and machine** detection performance and decision rationales

Dataset Creation

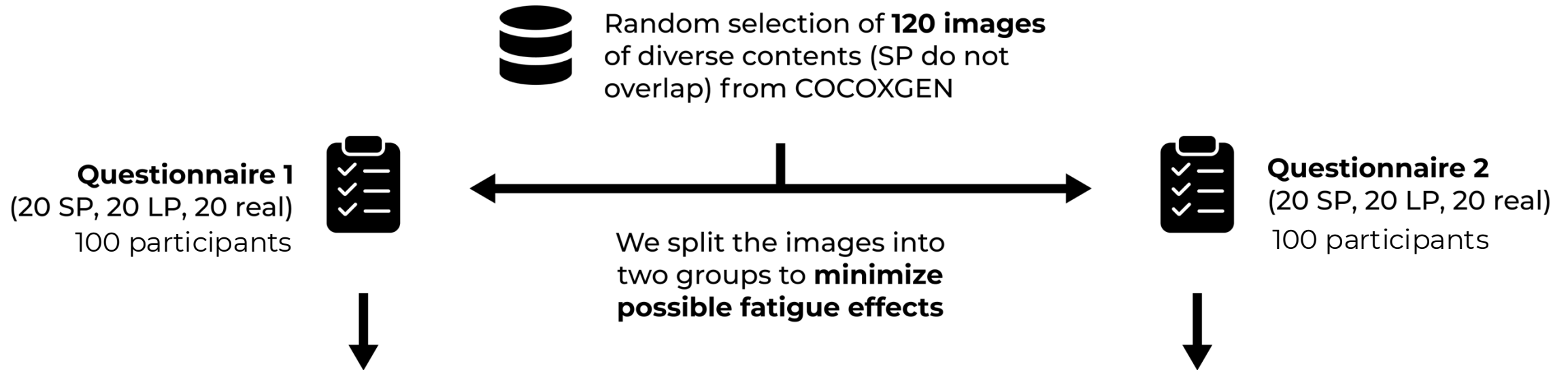


real	AI-generated			
1061	4244			
	LP		SP	
	2122		2122	
	Fooocus	SDXL	Fooocus	SDXL
	1061	1061	1061	1061


Image Generators:

- Stable Diffusion XL
- Fooocus (Juggernaut XL V8 + Realistic Vision V6)

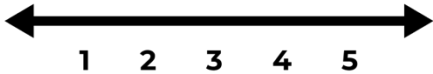
User Study



Real or fake?



How certain are you?



Any specific areas which influenced your decision?

- Yes, I will put it below.
- No it was a general impression.
- I am unsure.

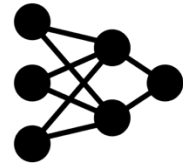
Which areas influenced your decision?

A1	B1	C1
A2	B2	C2
A3	B3	C3

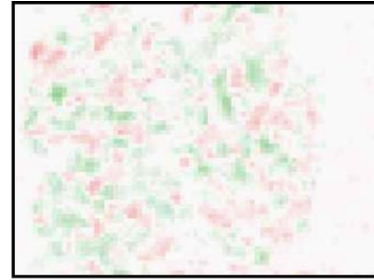
Model Evaluation



Original image



Grag2021
(Corvi et al., 2023)



Feature map
(60x80)



σ



< 0.5



real

≥ 0.5



fake

Classification
decision



Negative logit sum
per area



Positive logit sum
per area

Construction of two **heatmaps**: separate summation of negative and positive logits

Average
all values

Apply
sigmoid

Comparison of Detection Performance

Average performance of all study participants

Subset	Positives	F1
All	Real	0.7793
All	AI	0.8583
SP	AI	0.8002
LP	AI	0.8697
Foocus	AI	0.7857
SDXL	AI	0.8822
COCO	Real	0.7793

Statistically significant
(Wilcoxon test)

Grag2021 performance

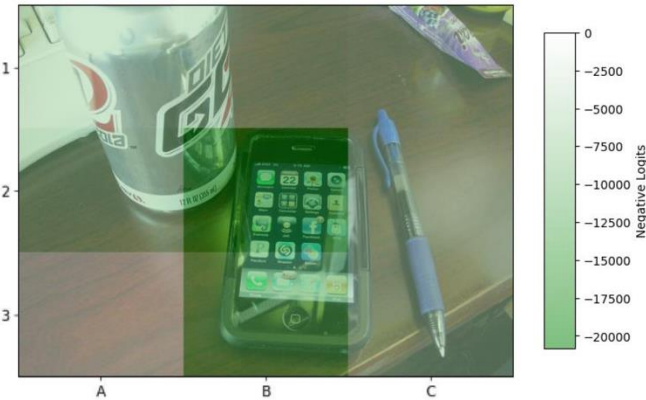
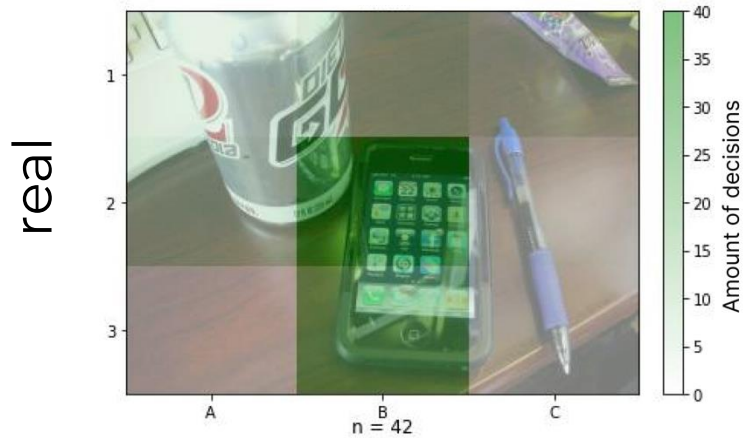
Subset	Positives	F1
All	Real	0.6957
All	AI	0.7200
SP	AI	0.7097
LP	AI	0.7302
Foocus	AI	1.0000
SDXL	AI	0.2222
COCO	Real	0.6957

Not statistically significant
(permutation test)

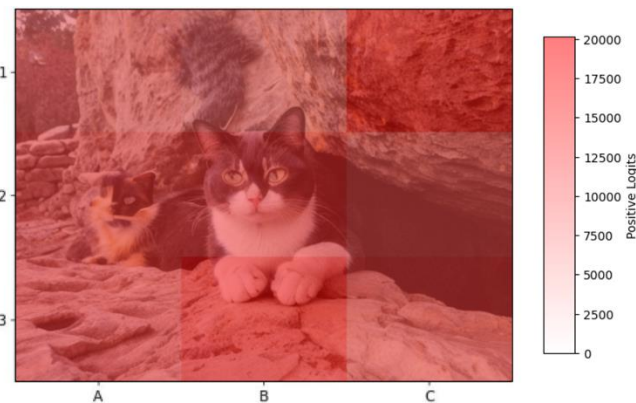
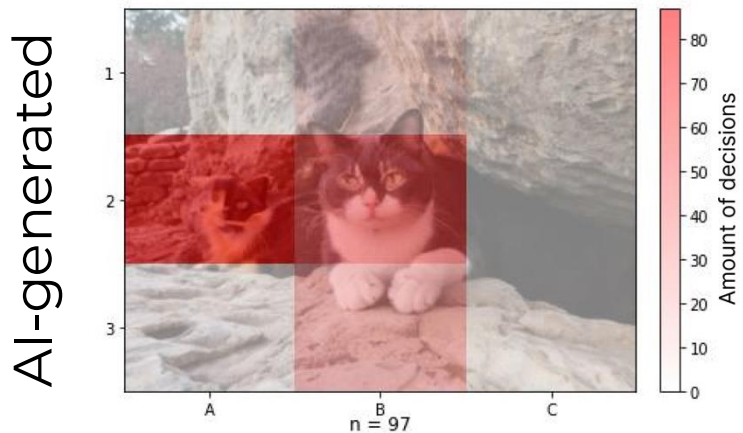
Comparison of Decision Rationales

Participants' heat map

Model heat map



People tend to select **clearly separable objects**; majority decides based on **“general impression”**



Detector: has high activation for **fine details** for most of the **AI-generated** images + for **clear object edges** for most of the **real** images

Findings

- **Machine detector performed worse** in distinguishing real from AI-generated images
 - **Due to** partially **downsampling** AI-generated images
 - **Weakens the practical applicability** of current machine detectors
- **Detection performance** for images generated with **long and detailed prompts** is **significantly higher** than for images generated with short less detailed prompts [L] [SEP]
 - Generator has to deviate more from its training data to fulfill the needs of a complex prompt
- **Visible artifacts do not need to be present** in the image for people **to be skeptical** about the authenticity of an image [L] [SEP]
 - **Opportunity for** the **human** ability to detect AI-generated images

Conclusion



Paper



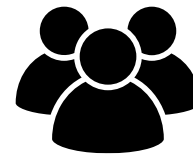
COCOXGEN

1. **Does the prompt's level of detail have an impact on detection performance?**
Image from more detailed prompts can be detected more easily as fake (by humans and machine detector).
2. **Does human or AI-model performance dominate?**
The average human had a better detection performance (due to downsampling).
3. **Do they consider the same objects and structures in an image when evaluating it?**
We could not find any significant overlap of decision rationales.

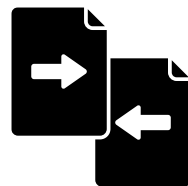


COCOXGEN

novel benchmark dataset



Conduction of a
large-scale user study



Direct **comparison of human and machine** detection performance and decision rationales

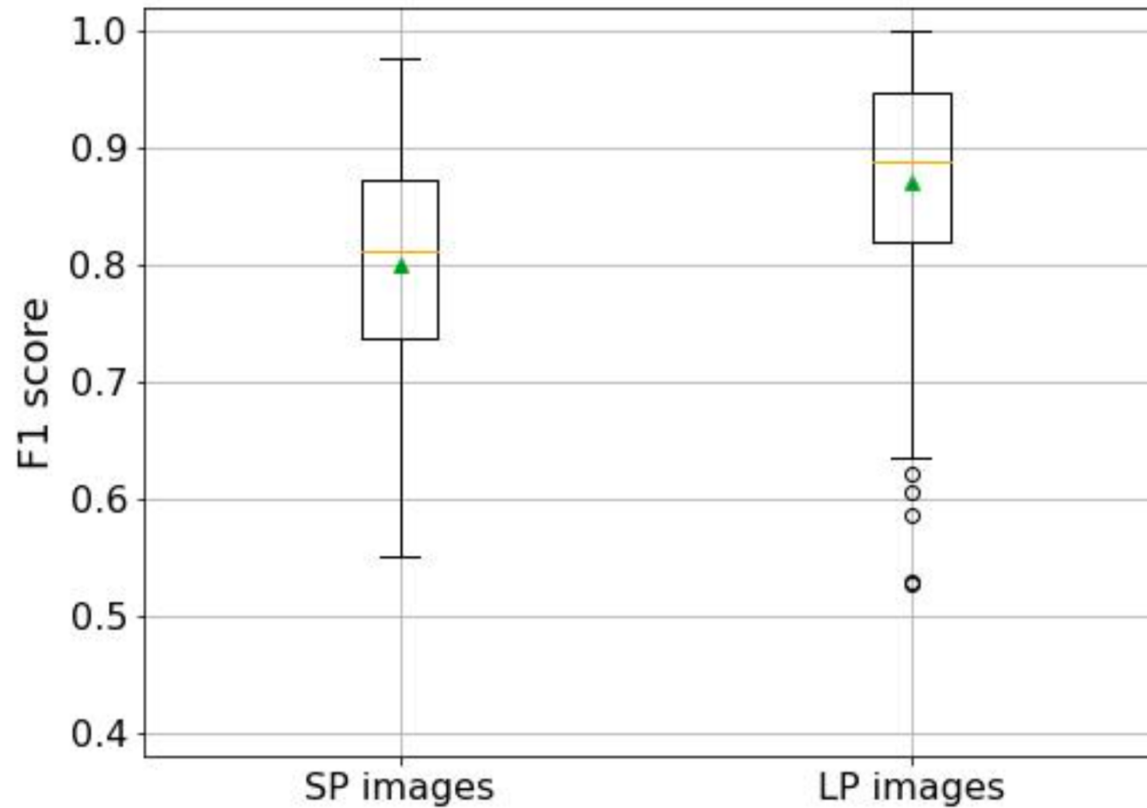
Decision Rationales

Percentage of participants' decisions based on "general impression" or selection of image areas

	"general impression"	selection of areas
all	48.78 %	45.16 %
AI-generated images	40.56 %	54.40 %
correctly classified as AI	31.78 %	65.28 %
incorrectly classified as real	74.71 %	10.69 %
real images	65.23 %	29.98 %
correctly classified as real	71.45 %	26.60 %
incorrectly classified as AI	33.33 %	63.31 %

User Study – Results

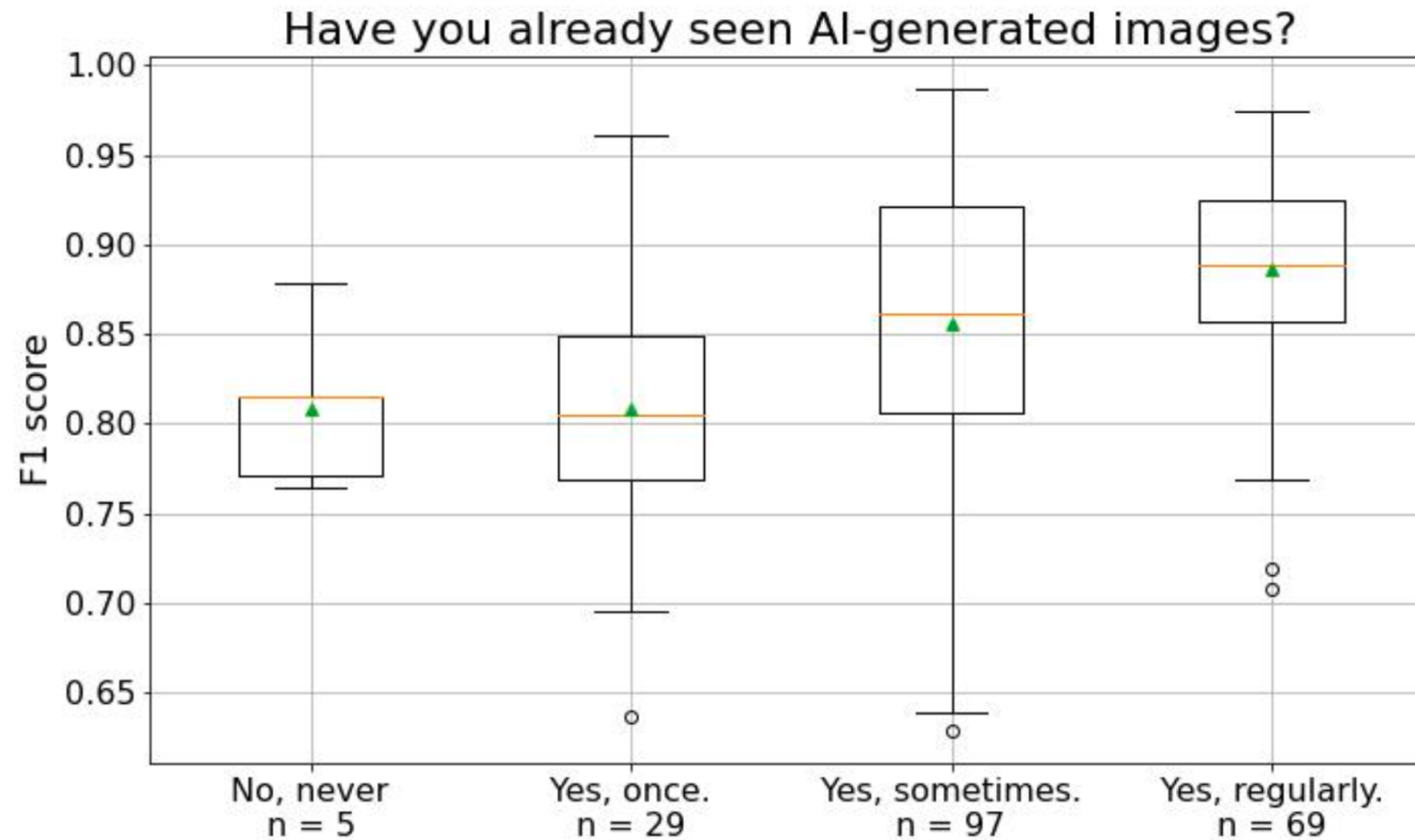
F1 scores of all study participants for SP and LP images



Wilcoxon test:
 $p = 2.1696e^{-22}$

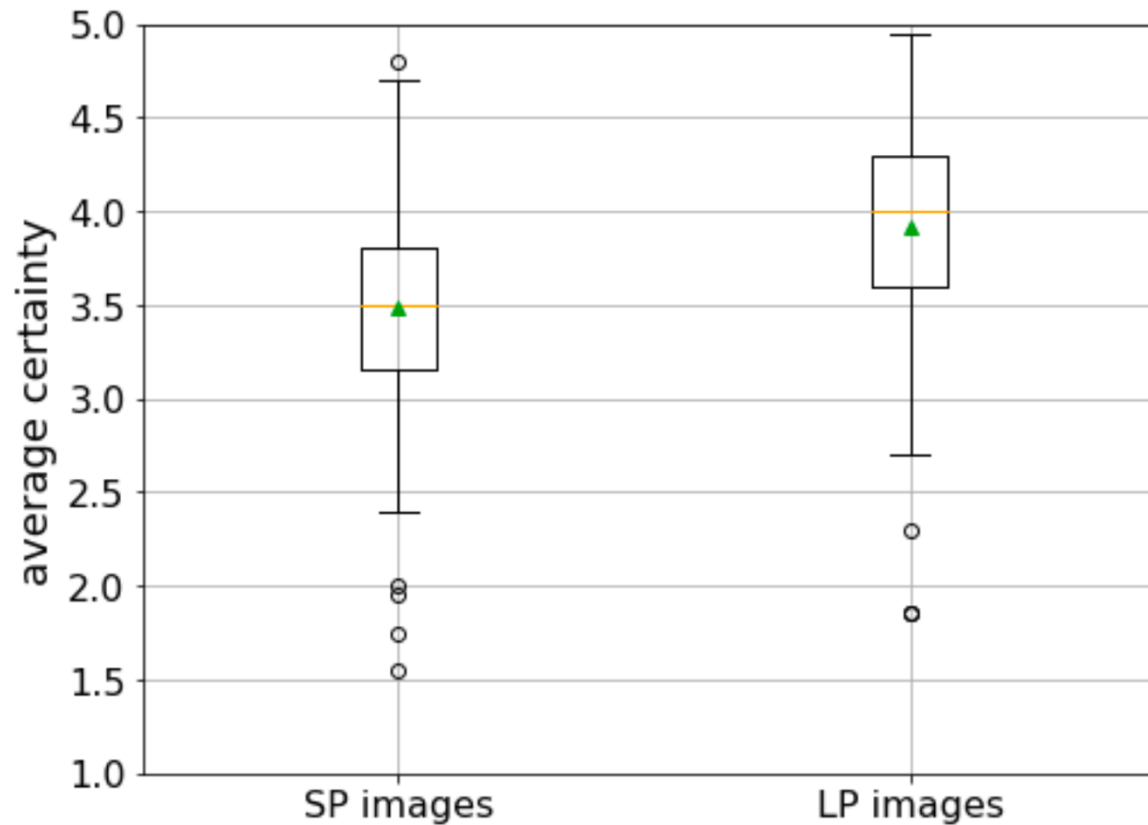
User Study – Results

F1 scores of all study participants of different experience levels



User Study – Results

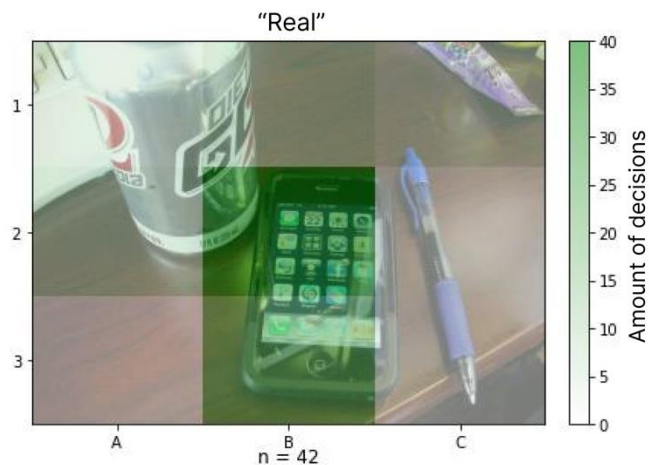
Average decision certainty of all study participants for SP and LP images



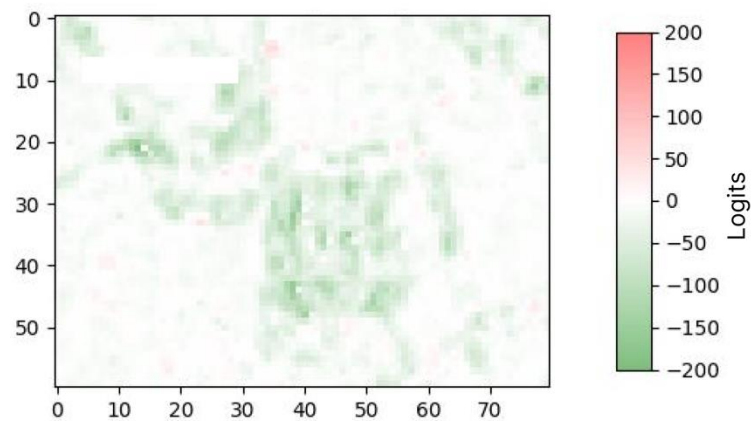
Wilcoxon test:
 $p = 1.2486e^{-29}$

Comparison of Decision Rationales

Participants' heat map



Model feature map



Model heat map



"AI-generated"

