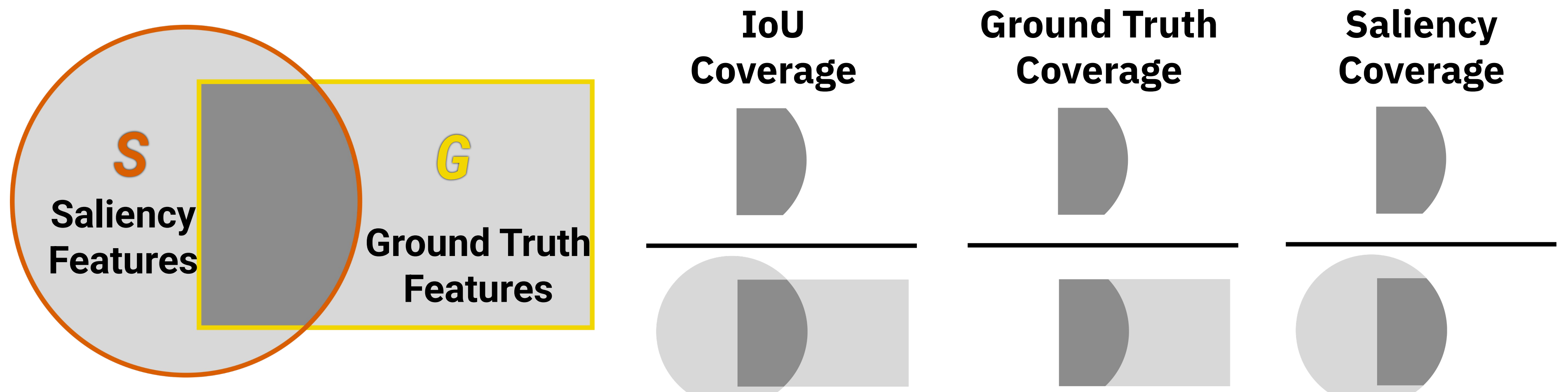
Shared Interest: Large-Scale Visual Analysis of Model Behavior by Measuring Human-AI Alignment

Angie Boggust[1], Benjamin Hoover[2], Arvind Satyanarayan[1], Hendrik Strobelt[2,1]

[1]MIT CSAIL, [2]IBM Research

Shared Interest Metrics: Computing Human and Model Agreement

To compute human-AI agreement on a data instance, we compute three complementary metrics between the saliency and ground truth features: IoU, Ground Truth, and Saliency Coverage.



Low Scores: saliency is disjoint from the ground truth.

High IoU Coverage: saliency and ground truth features are identical.

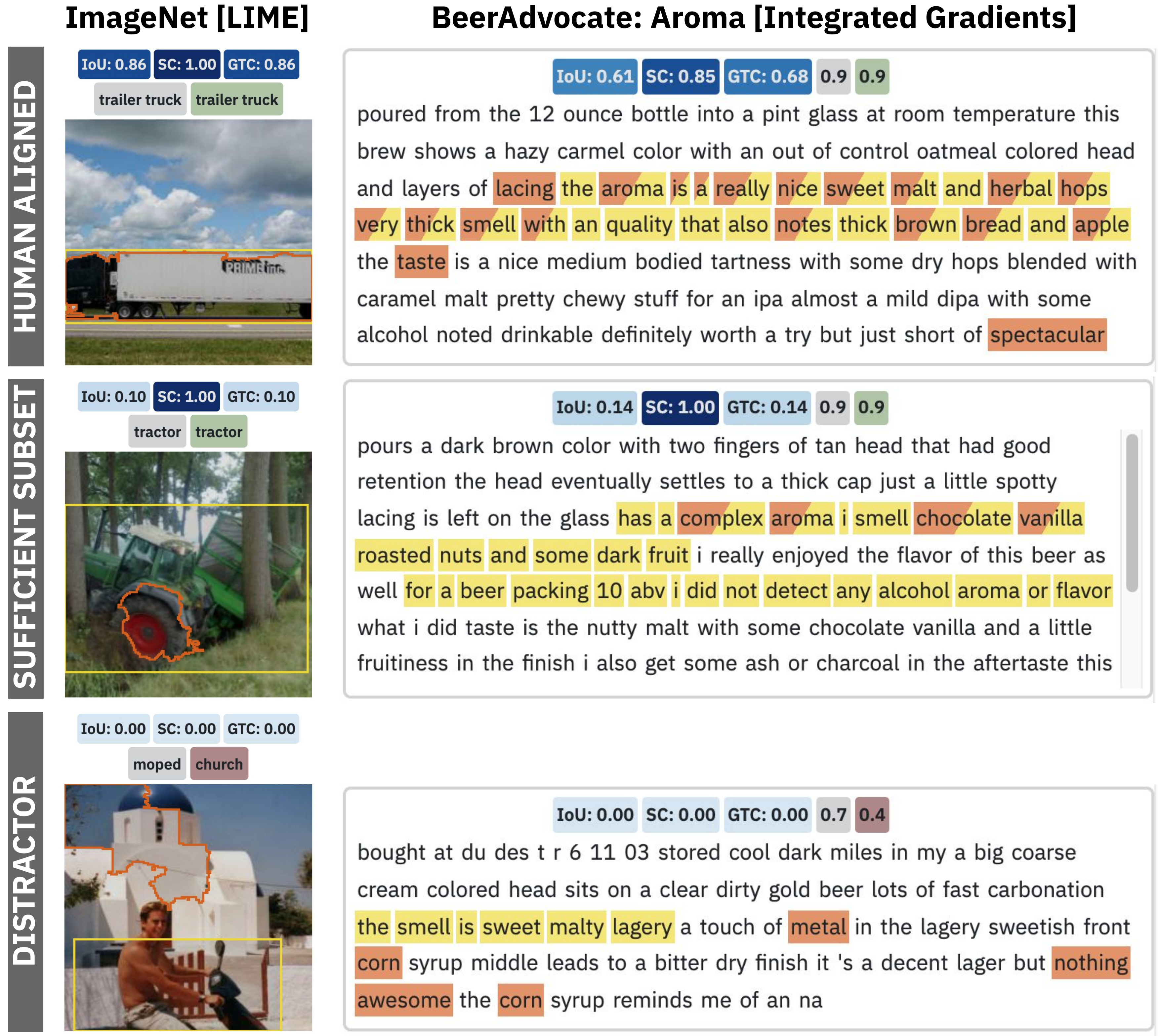
High Ground Truth Coverage: all ground truth features are salient to the model.

High Saliency Coverage: only ground truth features are salient to the model.

Identifying Recurring Patterns in Model Behavior

Computing Shared Interest metrics for every instance in a dataset enables us to sort, rank, and aggregate based on model behavior.

We surface common cases that identify dataset limitations and suggest avenues for future model iterations:


ImageNet [LIME]

BeerAdvocate: Aroma [Integrated Gradients]

poured from the 12 ounce bottle into a pint glass at room temperature this brew shows a hazy carmel color with an out of control oatmeal colored head and layers of lacing the aroma is a really nice sweet malt and herbal hops very thick smell with an quality that also notes thick brown bread and apple the taste is a nice medium bodied tartness with some dry hops blended with caramel malt pretty chewy stuff for an ipa almost a mild dipa with some alcohol noted drinkable definitely worth a try but just short of spectacular

pours a dark brown color with two fingers of tan head that had good retention the head eventually settles to a thick cap just a little spotty lacing is left on the glass has a complex aroma i smell chocolate vanilla roasted nuts and some dark fruit i really enjoyed the flavor of this beer as well for a beer packing 10 abv i did not detect any alcohol aroma or flavor what i did taste is the nutty malt with some chocolate vanilla and a little fruitiness in the finish i also get some ash or charcoal in the aftertaste this

bought at du des t r 6 11 03 stored cool dark miles in my a big coarse cream colored head sits on a clear dirty gold beer lots of fast carbonation the smell is sweet malty lagery a touch of metal in the lagery sweetish front corn syrup middle leads to a bitter dry finish it 's a decent lager but nothing awesome the corn syrup reminds me of an na

Interactive Probing to Understand Learned Concepts

Shared Interest can also be used as a mechanism to query model behavior.

Given a human annotation, we identify what the model 'knows' about that region by identifying the classes with the highest Shared Interest scores.

Using this procedure, we find our model has not only learned the concept of *dog* (the true label), but also relates the hat annotation to *sombrero*.