LERF: Language Embedded Radiance Fields
Appendix

A. Videos

We provide videos illustrating queries in 9 scenes on our project website. In all videos, the relevancy map images are post-processed such that all pixel values with relevancy score less than 0.5 are fully transparent. We choose this threshold because relevancy values under 0.5 means the rendered language embedding is more similar to the canonical negative phrases used ("object", "things", "stuff", "texture") than to the positive query prompt, so we consider them irrelevant to the query. This relevancy map is overlaid on the RGB renders. Relevancy map scale selection and normalization factor is constant across the entire video sequence to show scene-wide 3D consistency. During post-process editing we select queries which are visible during the specific segment of video; but we also provide full raw, uncut videos of outputs for the queries in the kitchen scene to give readers an idea of the scene-wide activations. Many other scenes contain wide-angle shots to observe relevancy maps scene-wide.

One notable property of relevancy maps across the scene that is more apparent in videos is that regions similar to, but not matching the query are also assigned a non-zero relevancy score, though not as high. For example, the query for *"utensils"* highlights most on the utensils in the dish drainer, but also on the knives hanging on the wall and utensils in the sink. The query for *"wooden spoon"* is most activated on the spoon, but also activates on other wooden components of the scene. This could be viewed as either a positive aspect of the language field in that it naturally groups similar regions to a query together, or a downside in that it can provide too many relevant regions in addition to the highest activation. For example, for the *"refrigerator"* query, the highest activation is assigned to the refrigerator, but much of the remaining kitchen space is also labeled as relevant. We hypothesize that this is related to the lack of grounding with visual-language models like CLIP. We find that the longer-tail and more specific the query, the more separation it tends to have with canonical phrases and hence the result more obviously pops out from other objects. Another effect visible in videos is the presence of "floaters" in the scene which can produce spurious activations, as discussed at length in Fig.5 and Sec.F.

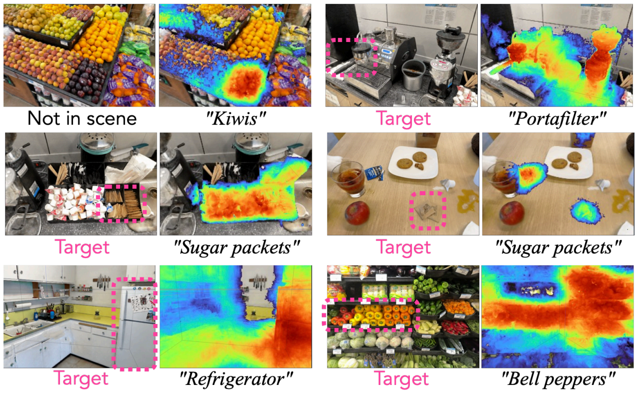


Figure 1. **Language and visual ambiguities from CLIP**: Cases with incorrect relevancy renders. Some failures can be attributed to visual similarity to the query (eg *"bell peppers"* gets distracted on jalepenos, *"portafilter"* activates on the grinder spout which has a similar metal cylindrical appearance, and *"refrigerator"* slightly activates on the white rectangular cabinets). Others are more flat-out failures, with *"sugar packets"* seemingly a confusing case to detect, and *"kiwis"* activating strongly on plums rather than correctly predicting nothing.

B. Additional qualitative results

We present a more complete list of results from scenes not pictured in the main text or videos in Fig. 7.

C. Numerical relevancy scores

We explore the reliability of using relevancy as a threshold for existence determination in and report precision-recall curves. Here, we provide raw relevancy scores for the queries in Fig. 1 of the main text to illustrate the behavior across different types of queries. Scores are shown in Table 1. One can observe that highly descriptive queries including visual and semantic properties produce higher relevancy scores (eg *"blue dish soap"*), and the lowest are abstract queries like *"electricity"* or small objects in clutter with not many close-up views (*"wooden spoon"*).

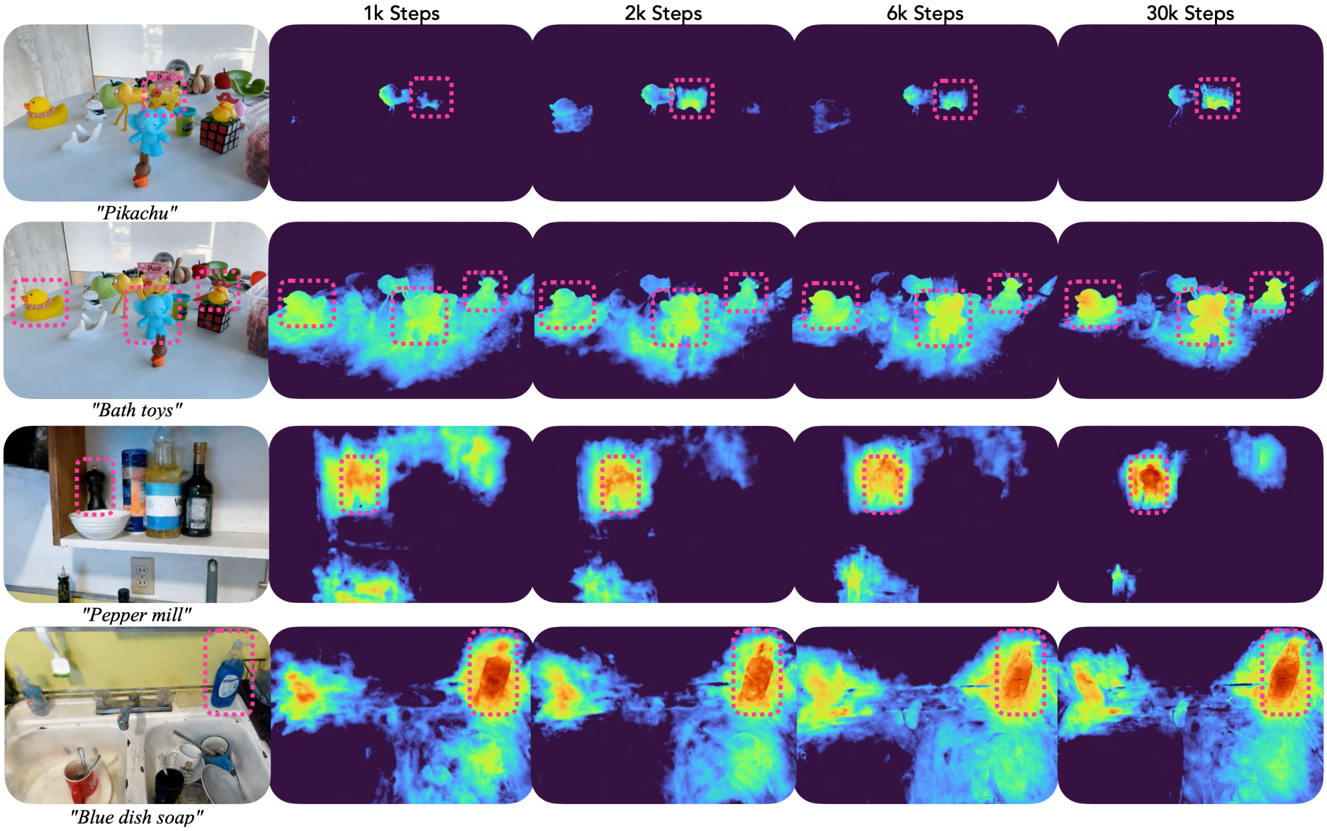


Figure 2. **LERF Convergence**. We visualize rendered relevancy maps at 1k, 2k, 6k, and 30k optimization steps. Relatively speaking, regions with more common semantics (like *"blue dish soap"*) and more expansive multi-view converge faster, while more fine-grained properties (like *"bath toys"*) take more steps. Notably, the optimization is quite stable: all queries produce reasonable activations within the first few thousand steps, and relevancy continues to refine over time.

D. Convergence speed

Videos and images for LERF were rendered after 30k optimization steps. However, usable relevancy maps can be obtained much sooner, as this section explores. We visualize relevancy maps and RGB renders of the kitchen scene and figurines scene after 1k, 2k, 6k, and 30k steps in Fig. 2. Because density converges significantly in NeRF within the first thousand steps, language embeddings in 3D are already usable at this stage. However, some fine-grained queries or small objects suffer in performance until later in optimization. In the *"Pikachu"* query, early training steps confuse the yellow figurine (Jake from Adventure Time) which is visually similar. As LERF converges, the actual Pikachu in the scene has higher relevancy than Jake. For fine-grained properties like *"bath toys"*, the relevancy starts out more blurry and becomes sharper and more isolated to the correct objects over time. Objects without much geometric separation like *"pepper mill"* also take longer to converge, since the surrounding geometry can be less precise.

Text query	Maximum relevancy score
utensils	0.77
wooden spoon	0.60
blue dish soap	0.83
waldo	0.76
paper towel roll	0.75
electricity	0.63
yellow	0.73
boops	0.77

Table 1. Maximum relevancy scores for each text query in Fig. 1 of main text, calculated from the displayed viewpoint. Highly specific queries have a higher relevancy value (*"blue dish soap"*, 0.83), while abstract queries can have lower ones (*"electricity"*, 0.63).

E. Experiment details

We provide a list of the labels used for the localization experiment in Tab. 2. Each label was labeled in 3-4 different views. We provide an exhaustive list of our custom long-tail labels for the existence experiment in Tab. 3.

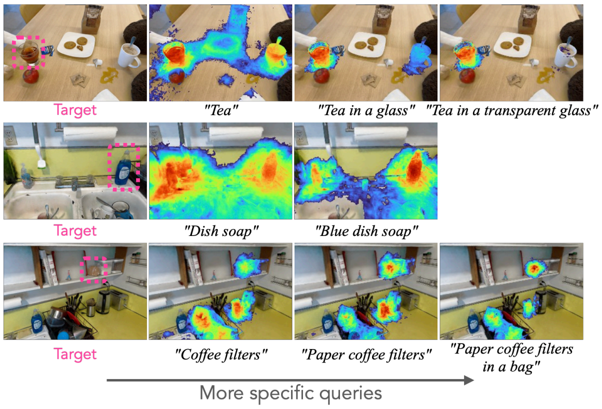


Figure 3. **Prompt tuning case study**: Some objects are sensitive to the prompt, with more specific wordings producing better results.

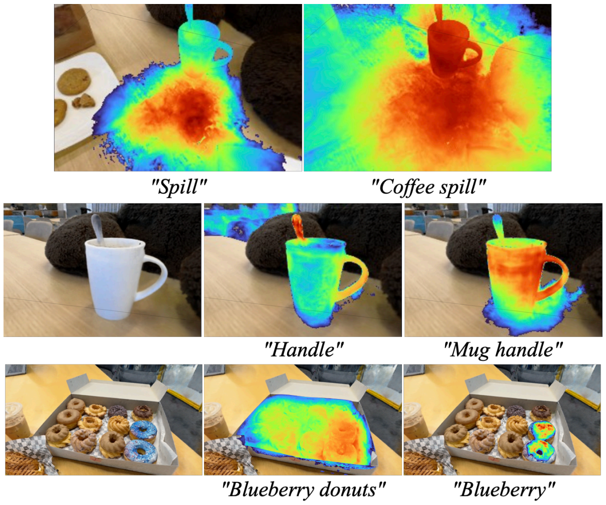


Figure 4. **CLIP bag-of-words behavior**: CLIP sometimes behaves as a bag-of-words, resulting in some adjectives not properly incorporating into queries.

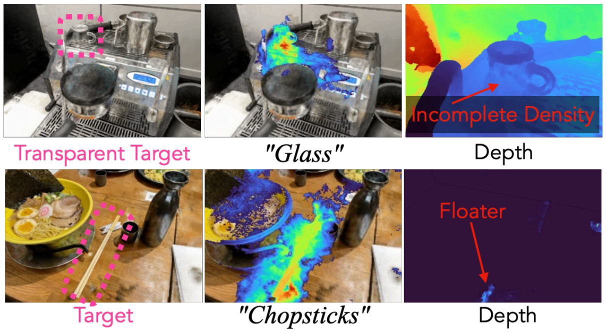


Figure 5. **Degradation with poor NeRF geometry**: Floaters and incomplete geometry can produce unreliable rendered CLIP embeddings.

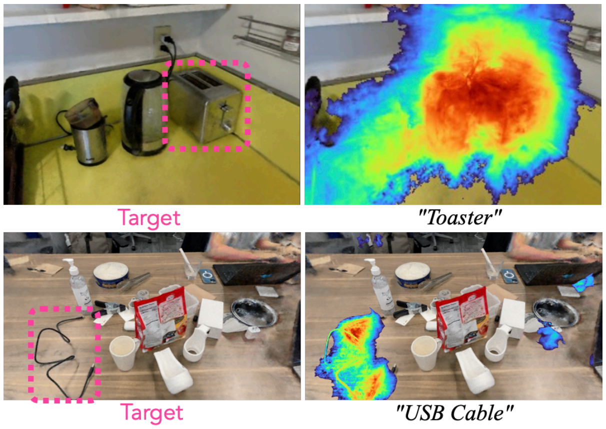


Figure 6. **Geometric separation impacts quality**: Queries without much geometric separation can blur between objects and foreground-background. In the toaster case, very few viewing angles were taken because of its position, which results in a fuzzier boundary.

F. Detailed Illustrations of Limitations

LERF inherits limitations from CLIP relating to language ambiguity and prompt sensitivity, as well as from NeRF's geometry representation capabilities. We present additional figures on failure cases to complement the ones provided in the main text.

Fig.1 showcases visual and language ambiguity from our usage of CLIP. Some queries get confused by unrelated regions of the scene because they appear very similar, such as the portafilter and the coffee grinder. In the refrigerator query, unrelated parts of the kitchen also activate in relevancy maps, though less strongly than refrigerator, because the CLIP embeddings of square white cabinets are more similar to a refrigerator than the canonical phrases. Sugar packets appear to be a confusing case for LERF, getting distracted in two separate scenes (teatime, espresso machine) with a tea packet and creamer pods respectively.

Fig.4 highlights another well-known undesirable property inherited from CLIP: text embeddings often behave as a bag-of-words rather than a grammatically parsed sentence. As a result, sometimes adding additional adjectives cause the output to latch onto incorrect regions (*"mug handle"* vs *"handle"* or *"coffee spill"* vs *"spill"*.)

Fig.5 shows performance degradation when geometry is unreliable in the underlying NeRF: reflective objects like the table in the ramen scene can produce holes, which result in CLIP embeddings from multiple views incorrectly averaging. Highly transparent objects like the glass cup in the espresso scene also suffer from lack of density, since the rendering weights mostly focus on the opaque background rather than the transparent foreground.

Fig.3 illustrates examples of relevancy maps improving in quality with subtly changed queries to become more and more specific. Usually this has a subtle effect on relevancy maps by refining the activation to a more localized region, for example providing progressively more descriptive queries improves the relevancy activation on the tea cup. This effect can also be drastic, for example *“Dish soap”* primarily activates on a pump soap bottle, but describing *“blue dish soap”* shifts focus to the correct object.

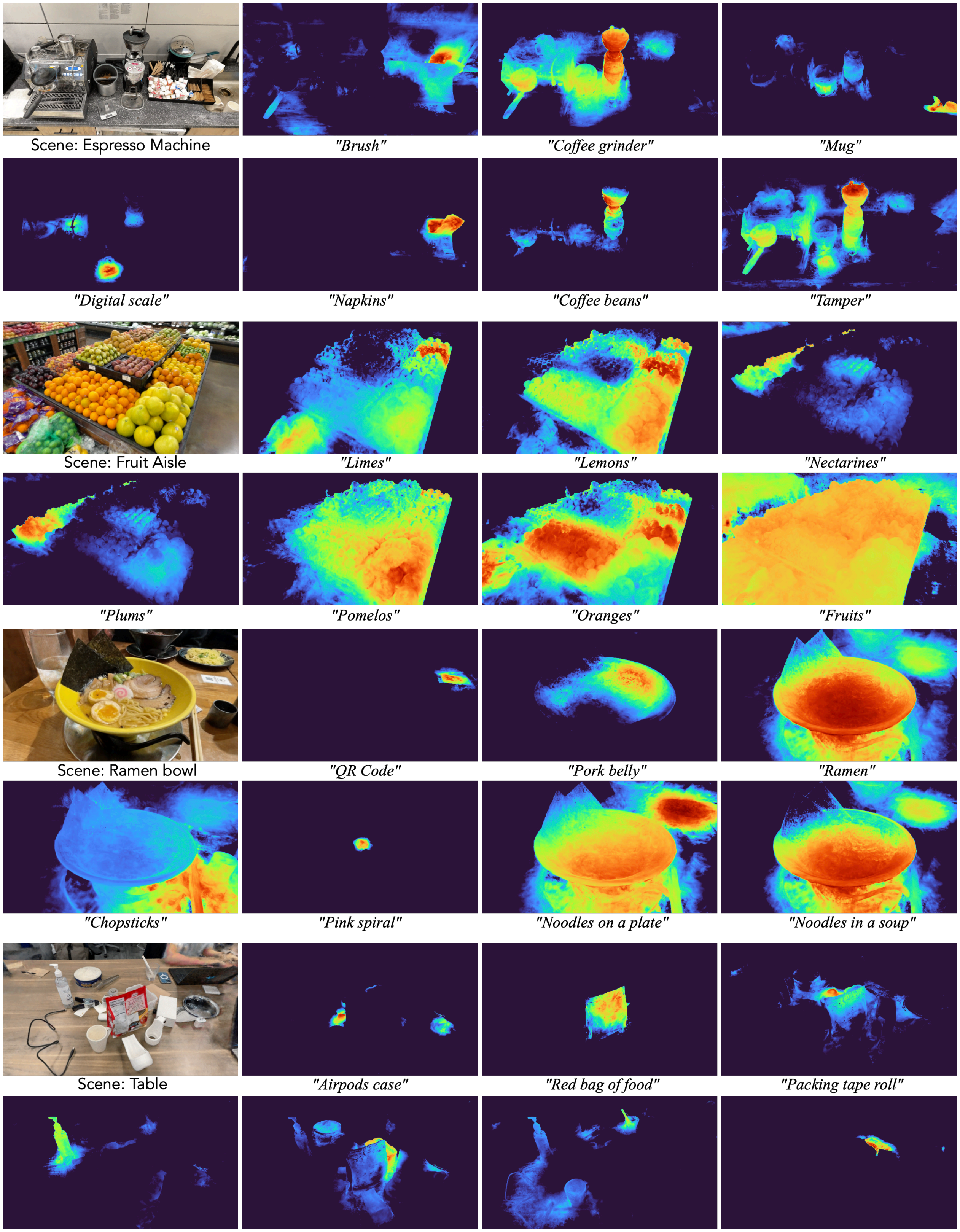
Finally, Fig.6 shows cases where lack of geometric separation (a cable close to a table, or a toaster flush in the corner) causes the relevancy maps to blur into other surrounding objects because most views of the background contain the foreground object in front.

Scene	Text queries			
Kitchen	blue hydroflask	coffee grinder	cookbooks	cooking tongs
	copper-bottom pot	dish soap	faucet	knives
	olive oil	paper towel roll	pepper mill	pour-over vessel
	power outlet	red mug	scrub brush	sink
	spice rack	utensils	vegetable oil	waldo
Bouquet	big white crinkly flower	bouquet	carnation	daisy
	eucalyptus	lily	rosemary	small white flowers
	vase			
Figurines	green apple	ice cream cone	jake	miffy
	old camera	pikachu	pink ice cream	porcelain hand
	quilted pumpkin	rabbit	red apple	rubber duck
	rubics cube	spatula	tesla door handle	toy cat statue
	toy chair	toy elephant	twizzlers	waldo
Ramen	bowl	broth	chopsticks	egg
	glass of water	green onion	napkin	nori
	pork belly	ramen	sake cup	wavy noodles
Teatime	bag of cookies	bear nose	coffee	coffee mug
	cookies on a plate	dall-e	hooves	paper napkin
	plate	sheep	spill	spoon handle
	stuffed bear	tea in a glass	yellow pouf	

Table 2. Labels used during detection experiments (75 total).

Scene	Positive Labels
Figurines	jake, miffy, rabbit, bunny, old camera, toy elephant, twizzlers, quilted pumpkin, tesla door handle, porcelain hand, rubics cube, rubber duck, apple, ice cream cone, pink ice cream, toy cat statue, toy chair, waldo, spatula, pikachu,table,
Kitchen	red mug, pour-over vessel, olive oil, vegetable oil, cookbooks, waldo, dish soap, plates, sink, faucet, copper-bottom pot, utensils, knives, spice rack, coffee grinder,flour, blue hydroflask, pepper mill, paper towel roll, scrub brush, power outlet, cooking tongs, transparent tupperware, coffee mug, coffee, spoon handle,
Teatime	stuffed bear, sheep, bear nose, coffee mug, spill, tea in a glass, cookies on a plate, bag of cookies, dall-e, hooves,coffee, yellow pouf, spoon handle, paper napkin, plate, wood, bag of food, hand sanitizer, mug,
Table	wood texture, red bag of food, hand sanitizer bottle, airpods case, usb cable, brown paper napkins, transparent tupperware, colorful coaster, packing tape roll, hardware clamps, iphone, laptop, metal cooking tongs, mug, drinking straw, table, tea in a glass,
Bouquet	big white crinkly flower, bouquet, carnation, daisy, eucalyptus, lily, rosemary small white flowers, table, flowers, pink flowers, wood, sink

Table 3. Long-tail labels used for existence experiment. Each row shows the positive labels for a given scene. Negative labels consist of the positives for other scenes, re-labeled to match the ground truth.



Scene: Espresso Machine *"Brush"* *"Coffee grinder"* *"Mug"*

"Digital scale" *"Napkins"* *"Coffee beans"* *"Tamper"*

Scene: Fruit Aisle *"Limes"* *"Lemons"* *"Nectarines"*

"Plums" *"Pomelos"* *"Oranges"* *"Fruits"*

Scene: Ramen bowl *"QR Code"* *"Pork belly"* *"Ramen"*

"Chopsticks" *"Pink spiral"* *"Noodles on a plate"* *"Noodles in a soup"*

Scene: Table *"Airpods case"* *"Red bag of food"* *"Packing tape roll"*

"Hand sanitizer bottle" *"Transparent tupperware"* *"Drinking straw"* *"Brown paper napkins"*

Figure 7. Additional results of scenes not reported fully in the main text or rendered in videos.