Panoptic Narrative Grounding Supplementary Material

Cristina González[1] Nicolás Ayobi[1] Isabela Hernández[1] José Hernández[1]
Jordi Pont-Tuset[2] Pablo Arbeláez[1]

[1]Center for Research and Formation in Artificial Intelligence, Universidad de los Andes, Colombia

[2]Google Research, Switzerland

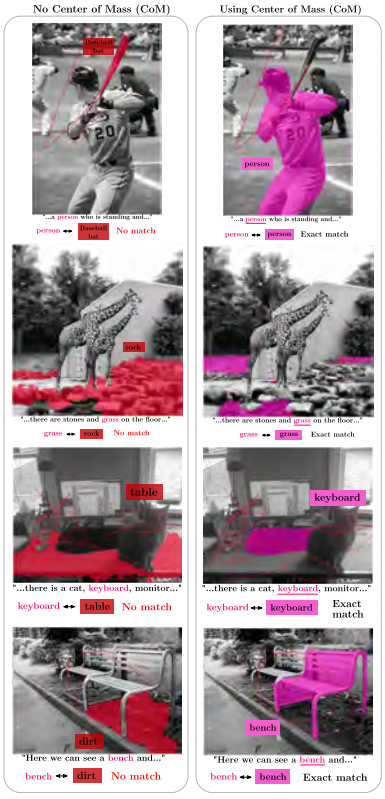


Figure 1: **Examples of Center of Mass (CoM) use.** Overall, averaging traces' spatial trajectories in Localized Narratives proves beneficial for noun phrase grounding purposes (pink colored). We provide annotations that have apt spatial accuracy, compared to annotations based on the temporal dimension, which consider trace times within segmentation regions (red colored). We summarize circling and irregular patterns, which are instinctive ways of pointing at objects in images and suitably capture the annotators' intention.

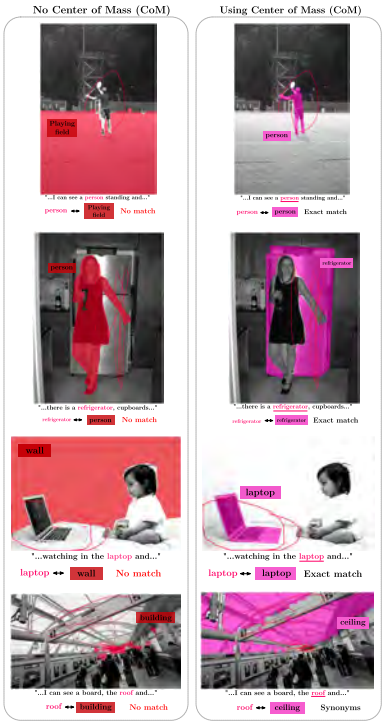


Figure 2: **Examples of Center of Mass (CoM) use.** Overall, averaging traces' spatial trajectories in Localized Narratives proves beneficial for noun phrase grounding purposes (pink colored). We provide annotations that have apt spatial accuracy, compared to annotations based on the temporal dimension, which consider trace times within segmentation regions (red colored). We summarize circling and irregular patterns, which are instinctive ways of pointing at objects in images and suitably capture the annotators' intention.

Figure 3: **Examples of textual *synonym*, *hierarchical* and *meronym* relationships.** Tag clouds showing category words (colored) used to refer MS COCO object categories (gray). First row depicts *synonym* relationships, second row depicts *hierarchical* relationships, while third row depicts *meronym* relationships between input nouns and object categories. Word size indicates frequency of appearance in captions.

Figure 4: **Example of vicinity region analysis.** The spoken noun phrase "skateboard" does not match any region when only considering the CoM region assignment (left). If we extend the candidate segments (right), we obtain a match with "skateboard", which counters the time and spatial shifts between the pronunciation of the words and pointing to the correct regions. Best viewed in color.

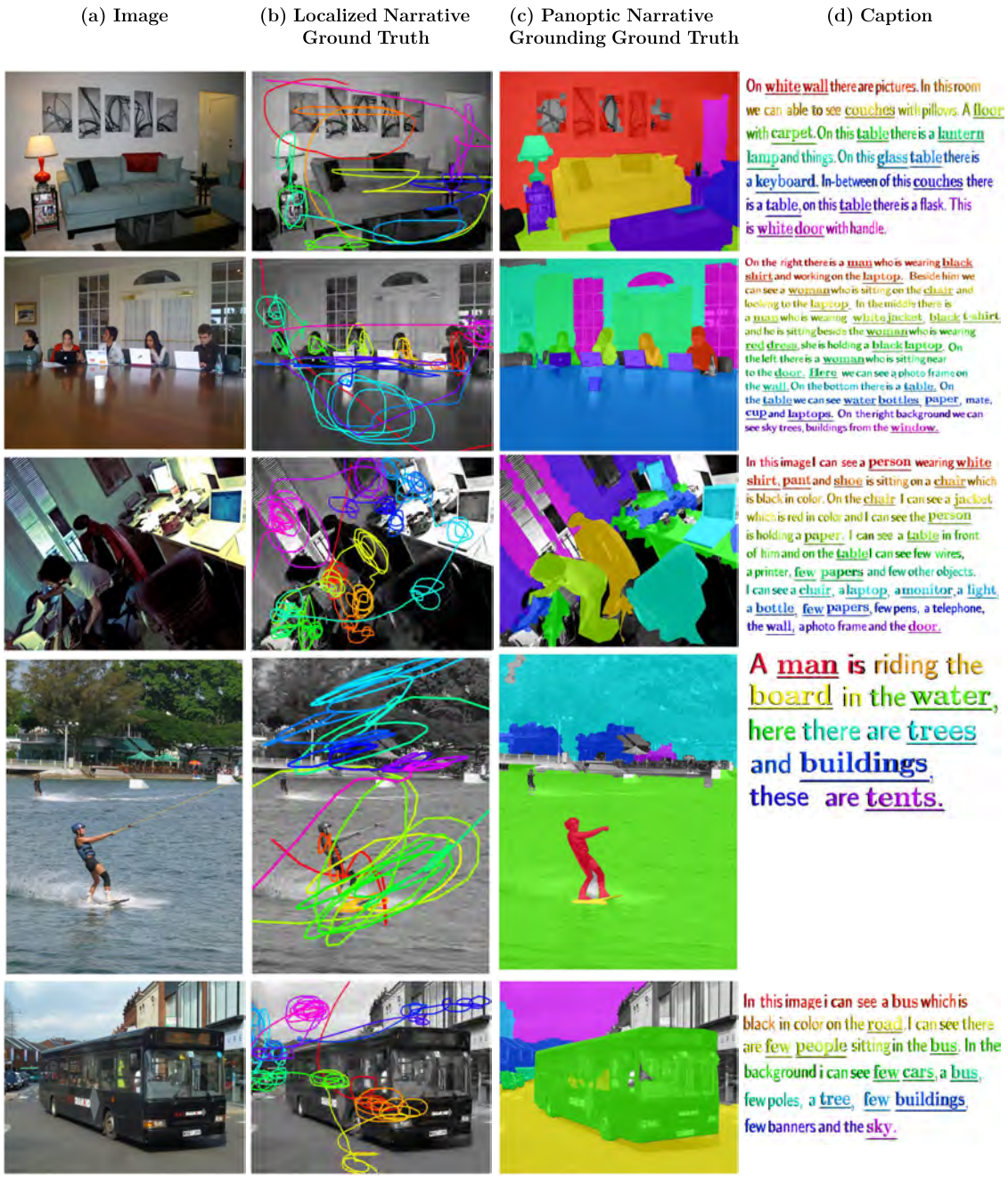
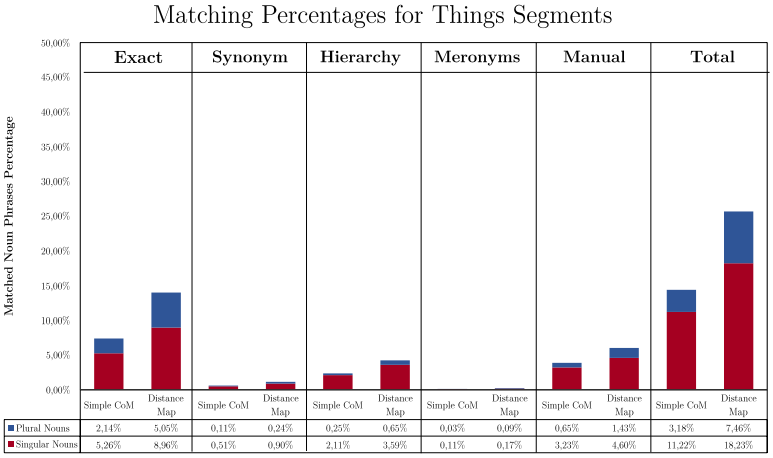
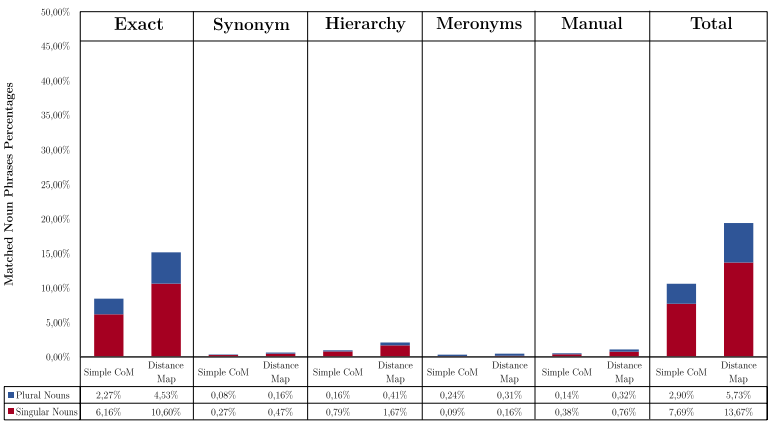


Figure 5: **Additional Ground-Truth Annotations Results.** Examples of different Panoptic Narrative Grounding ground truth resulting from the proposed annotation transfer algorithm (c). We show the input image (a) and Localized Narrative traces (b) and caption with the matched panoptic segmentation regions (d). Color gradient in the trace, panoptic segmentation and caption indicates time over the language. The segmentation regions are visualized with the color of their corresponding noun phrases, according to the last associated spoken word.



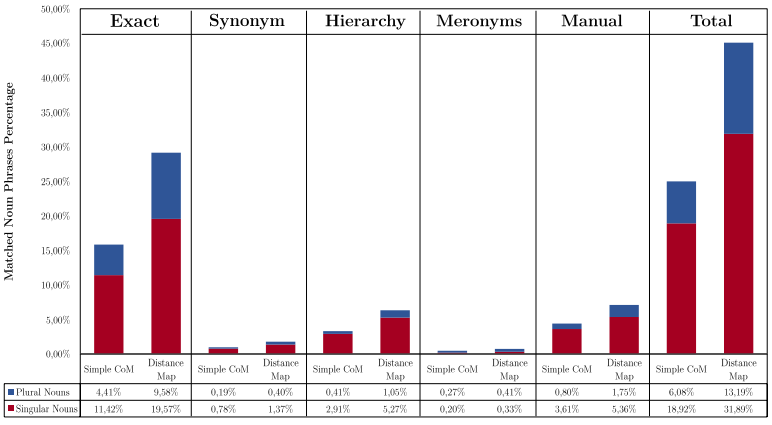
	Exact		Synonym		Hierarchy		Meronyms		Manual		Total	
	Simple CoM	Distance Map	Simple CoM	Distance Map	Simple CoM	Distance Map	Simple CoM	Distance Map	Simple CoM	Distance Map	Simple CoM	Distance Map
Plural Nouns	2,14%	5,05%	0,11%	0,24%	0,25%	0,65%	0,03%	0,09%	0,65%	1,43%	3,18%	7,46%
Singular Nouns	5,26%	8,96%	0,51%	0,90%	2,11%	3,59%	0,11%	0,17%	3,23%	4,60%	11,22%	18,23%

(a) Matching percentage statistics for only regions corresponding to things.



	Exact		Synonym		Hierarchy		Meronyms		Manual		Total	
	Simple CoM	Distance Map	Simple CoM	Distance Map	Simple CoM	Distance Map	Simple CoM	Distance Map	Simple CoM	Distance Map	Simple CoM	Distance Map
Plural Nouns	2,27%	4,53%	0,08%	0,16%	0,16%	0,41%	0,24%	0,31%	0,14%	0,32%	2,90%	5,73%
Singular Nouns	6,16%	10,60%	0,27%	0,47%	0,79%	1,67%	0,09%	0,16%	0,38%	0,76%	7,69%	13,67%

(b) Matching percentage statistics for regions corresponding to stuff.



	Exact		Synonym		Hierarchy		Meronyms		Manual		Total	
	Simple CoM	Distance Map	Simple CoM	Distance Map	Simple CoM	Distance Map	Simple CoM	Distance Map	Simple CoM	Distance Map	Simple CoM	Distance Map
Plural Nouns	4,41%	9,58%	0,19%	0,40%	0,41%	1,05%	0,27%	0,41%	0,80%	1,75%	6,08%	13,19%
Singular Nouns	11,42%	19,57%	0,78%	1,37%	2,91%	5,27%	0,20%	0,33%	3,61%	5,36%	18,92%	31,89%

(c) Matching percentage statistics for regions corresponding both things and stuff.

Figure 6: **Annotation transfer statistics.** Percentage statistics of nouns phrases matched with the category of the assigned region. We counted the amount of matched noun phrases for each type of match (Exact, Synonym, Hierarchical, Meronym or Manual) and with and without the inclusion of distance analysis (Distance Map).

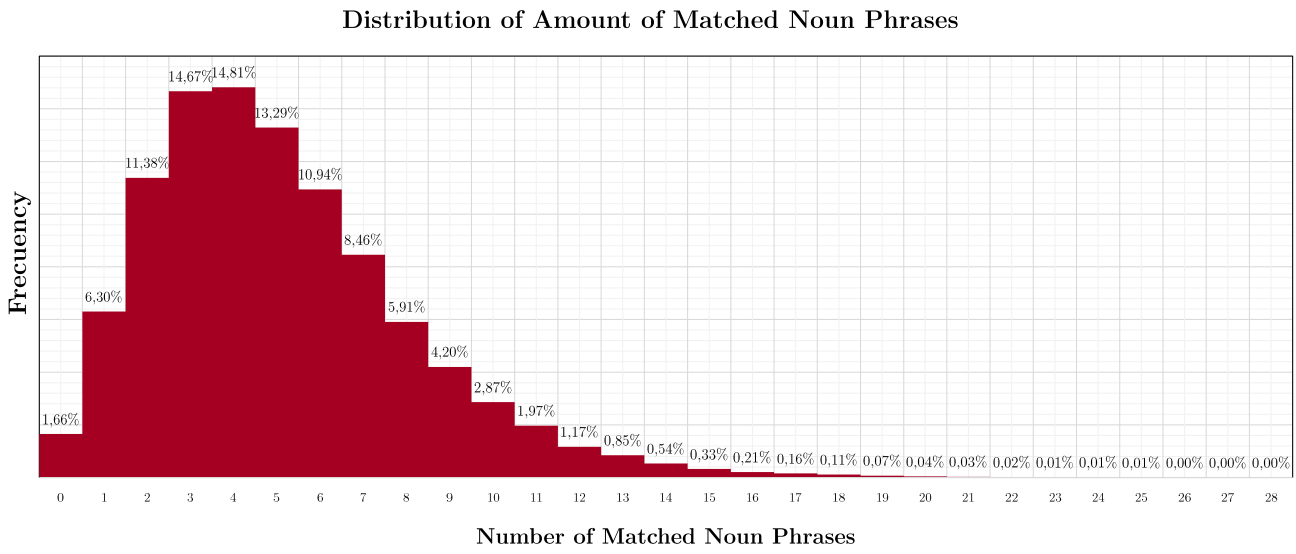


Figure 7: **Density histogram** of the amount of matched noun phrases per localized narratives in the entire dataset. We show the amount of matches distribution for the final algorithm that includes stuff, thing and takes into account the distances analysis.

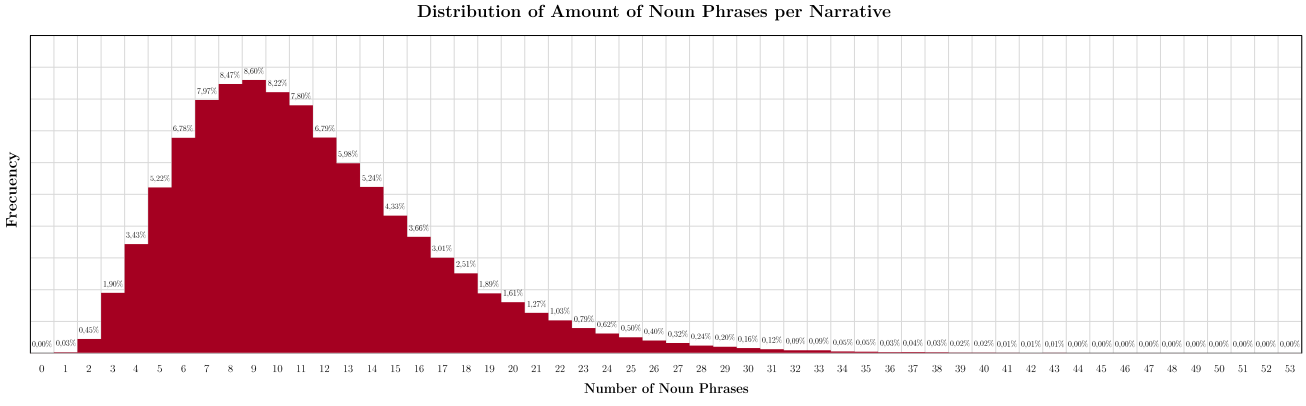


Figure 8: **Density histogram** of the amount of identified noun phrases per localized narratives in the entire dataset. We show the amount of noun phrases distribution for the final algorithm that includes stuff, thing and takes into account the distances analysis.

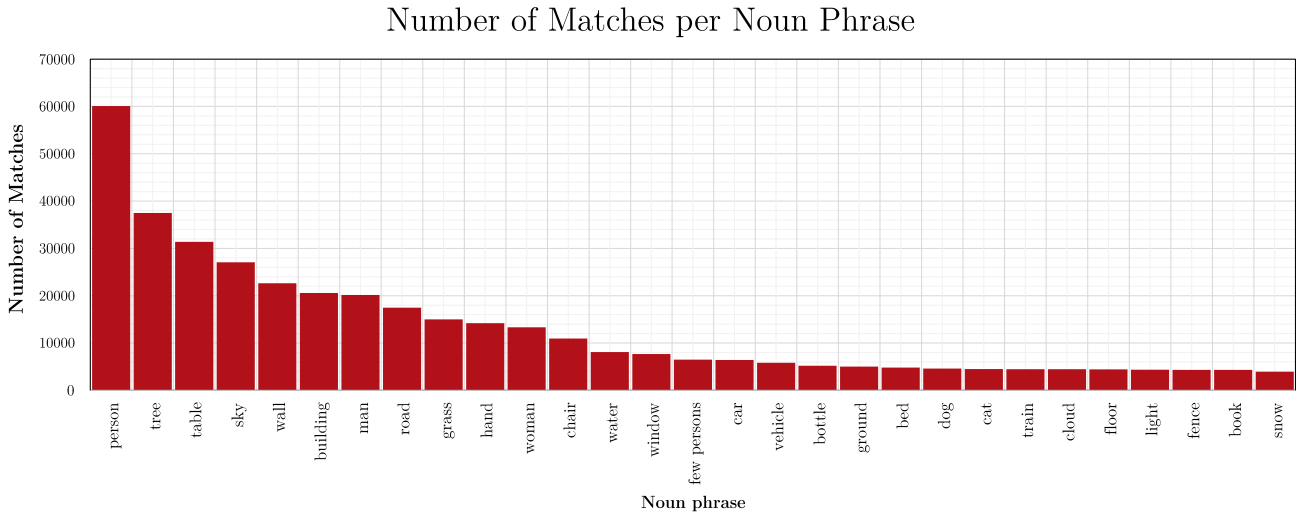


Figure 9: **Number of mathces histogram** for the 30 most common noun phrases in the dataset. We demonstrate that the frequency of matching of certain noun phrases is way higher than the rest, thus showing and evident long tail in the noun phrases distribution.

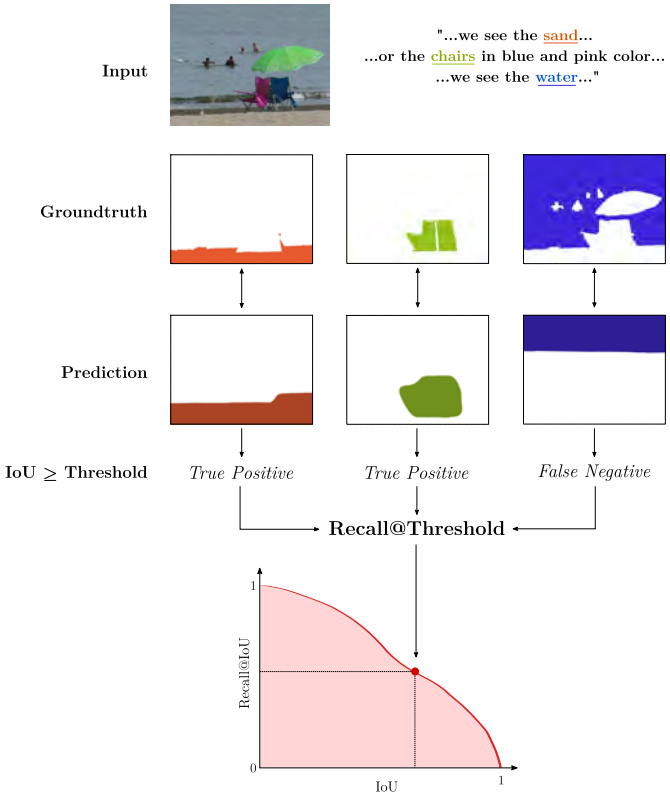


Figure 10: **Formulation for Panoptic Narrative Grounding evaluation through Average Recall.** For each image-caption pair, we compare all predictions to their corresponding ground-truth annotations via the IoU measure. Sequently, we use different thresholds in this measure (**IoU**) to determine particular sets of positive and negative detections and a recall value (**Recall@Threshold**). The area under the resulting curve is regarded as Average Recall and is proposed as an evaluation metric for the Panoptic Narrative Grounding task.

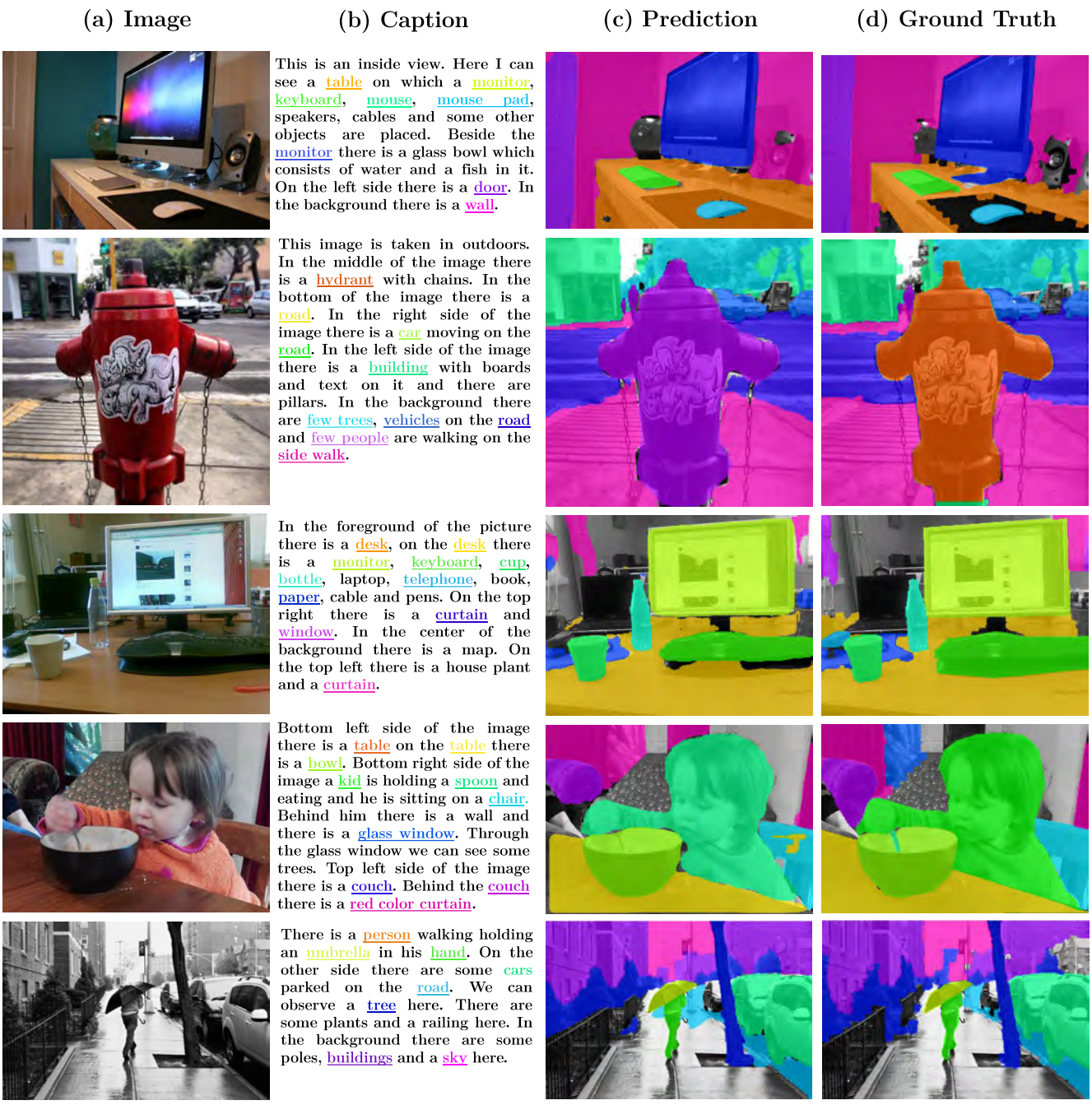


Figure 11: **Qualitative results for Panoptic Narrative Grounding.** Example predictions of our baseline in the validation split of our benchmark. The inputs are the image (a) and the caption without highlighted noun phrases (b). The outputs are a set of noun phrases in the caption (b), each with a corresponding region in the predicted segmentation (c). (d) shows the ground-truth panoptic segmentations. Best viewed in color.

Figure 12: **Qualitative results for Panoptic Narrative Grounding.** Example predictions of our baseline in the validation split of our benchmark. The inputs are the image (a) and the caption without highlighted noun phrases (b). The outputs are a set of noun phrases in the caption (b), each with a corresponding region in the predicted segmentation (c). (d) shows the ground-truth panoptic segmentations. Best viewed in color.

Figure 13: **Qualitative results for Panoptic Narrative Grounding.** Example predictions of our baseline in the validation split of our benchmark. The inputs are the image (a) and the caption without highlighted noun phrases (b). The outputs are a set of noun phrases in the caption (b), each with a corresponding region in the predicted segmentation (c). (d) shows the ground-truth panoptic segmentations. Best viewed in color.

Figure 14: **Qualitative results for Panoptic Narrative Grounding.** Example predictions of our baseline in the validation split of our benchmark. The inputs are the image (a) and the caption without highlighted noun phrases (b). The outputs are a set of noun phrases in the caption (b), each with a corresponding region in the predicted segmentation (c). (d) shows the ground-truth panoptic segmentations. Best viewed in color.