# **Panoptic Narrative Grounding Supplementary Material**

Cristina González<sup>1</sup>

Nicolás Ayobi<sup>1</sup> Jordi Pont-Tuset<sup>2</sup> Isabela Hernández<sup>1</sup> Pablo Arbeláez<sup>1</sup> José Hernández<sup>1</sup>

<sup>1</sup>Center for Research and Formation in Artificial Intelligence, Universidad de los Andes, Colombia <sup>2</sup>Google Research, Switzerland



No Center of Mass (CoM)
Using Center of Mass (CoM)

Image: Comparison of Comp

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.







Matching Percentages for Stuff Segments





(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).



Distribution of Amount of Matched Noun Phrases

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.



Distribution of Amount of Noun Phrases per Narrative

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.



# Number of Matches per Noun Phrase

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.



# (b) Caption

This is an inside view. Here I can see a <u>table</u> on which a <u>monitor</u>, <u>keyboard</u>, mouse, mouse speakers, cables and some other objects are placed. Beside the monitor there is a glass bowl which consists of water and a fish in it. On the left side there is a <u>door</u>. In the background there is a wall.

This image is taken in outdoors. In the middle of the image there is a <u>hydrant</u> with chains. In the bottom of the image there is a <u>coad</u>. In the right side of the image there is a <u>car</u> moving on the road. In the left side of the image there is a <u>building</u> with boards and text on it and there are pillars. In the background there are few trees, vehicles on the road and few people are walking on the <u>side walk</u>.

In the foreground of the picture there is a  $\underline{\operatorname{desk}}$ , on the  $\underline{\operatorname{desk}}$  there is a <u>monitor</u>, <u>keyboard</u>, <u>cup</u>, bottle, laptop, telephone, book, paper, cable and pens. On the top right there is a <u>curtain</u> and window. In the center of the background there is a map. On the top left there is a house plant and a <u>curtain</u>.

Bottom left side of the image there is a <u>table</u> on the <u>table</u> there is a bowl. Bottom right side of the image a kid is holding a spoon and eating and he is sitting on a <u>chair</u>. Behind him there is a wall and there is a <u>glass window</u>. Through the glass window we can see some trees. Top left side of the image there is a <u>couch</u>. Behind the <u>couch</u> there is a <u>red color curtain</u>.

There is a person walking holding an <u>umbrella</u> in his <u>hand</u>. On the other side there are some cars parked on the <u>road</u>. We can observe a tree here. There are some plants and a railing here. In the background there are some poles, <u>buildings</u> and a <u>sky</u> here.

# pad.



(c) Prediction









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.







(d) Ground Truth



# (b) Caption

This picture is clicked on <u>road</u>. There is bus in the foreground. There are <u>few people</u> in the bus. On bus there is text. To the extreme right there is a car. In the background there is building, a banner to the pole and  $\underline{sky}$ . To the top left there are leaves of <u>tree</u>.

olue sky. This is a bare tree. Here we can see small poles in the form of fence in a ground. We can see <u>one man</u> standing here. We can see a man standing and flying a colourful kite with white and red colour. We can see a <u>man</u> standing beside to him. This is a grass. In this image there is a dog on the <u>bed</u>, and on the <u>bed</u> there are <u>blankets</u> and <u>pillow</u>, bag and some clothes and there is a television on a table in the background. On the <u>television</u> there are objects and there is wall, and there are objects on the right side of the image.

# (c) Prediction

(d) Ground Truth







In this picture we observe a eroplane flying in the sky and there are two propellers attached to it. There are also two wings in the <u>aeroplane</u>. The aeroplane is white and blue in  $\operatorname{color}$ 

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.



## (b) Caption

As we can see in the image there is a <u>water</u> and <u>four people</u>. The <u>women</u> over here is standing on <u>blue color surf board</u> and this <u>man</u> is sitting on <u>blue color surf board</u> and on the left side this <u>man</u> is laying on white color surf board.

I can see in this image a <u>bridge</u> on the <u>water</u>. On the left side of the image I can see a <u>boat</u> is floating on the <u>water</u>, a group of <u>people</u> on the roadside are standing together, a pole, <u>street lights</u>, <u>trees</u> and <u>vehicles</u> on the <u>road</u> and on the right side of the image I can see tower and <u>sky</u>.

In this image I can see there are the three persons standing in the right side and a person holding a book and carrying a hand book in front of them there is a book kept on the and there are some boxes kept on the floor and back side of them there is a refrigerator visible and right side corner there is a valid and some objects attached to the wall and left side corner there is a table of there are the boxes kept on the table there are the boxes kept on the table there are there persons standing and back side of them there is a table on the table there is a book kept on the table there is a book will be there is a table on the table there is a book will be there is a table on the table there is a book will be there is a table on the table there is a book will be table there is a book will be there is a book will be table on the table.

In the picture there is a <u>bird</u> inside the restaurant it is walking on the floor, there is a dining <u>table</u>, behind the <u>table</u> there is a <u>person</u> serving the <u>food</u> in front of <u>person</u> there are large variety of <u>food</u> items placed on the <u>table</u>, the <u>table</u> is of yellow and orange color.



This is the picture taken in a room, there are a group of <u>people</u> sitting on <u>chairs</u> in front of these <u>people</u> there is a <u>table</u> and it is covered with a cloth and on the <u>table</u> there are tissues, cloth, plates, glasses, <u>bowl</u>, basket and some food items. Beside the <u>people</u> there is a wall with photo and on top of the wall there are <u>trees</u> with decorative lights.

## (c) Prediction

(d) Ground Truth



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.



### (b) Caption

In this image there is a <u>table</u> and on top of it there is a plate with <u>few doughnuts</u> on it. On top of <u>doughnuts</u> there is a <u>cream</u> and <u>chocolate</u> on it. This image is taken inside a room. (c) Prediction

(d) Ground Truth





A <u>giraffe</u> is trying to sit by the side of a <u>tree</u>. There is another <u>giraffe</u> coming from the behind. There are <u>trees</u> around them.



In the picture there are <u>many</u> <u>houses</u>, near to the <u>houses</u> there are many plants, there is a <u>small lake</u>, there is a <u>clock</u> <u>tower</u> near to the <u>house</u>, there are <u>small house plants</u> in the <u>houses</u>, there is a <u>dark sky</u>.



In this image in the middle there is a train and engine inside that there is person. On the left there is platform on that there are many people, house. On the left there is a man he wears t\_shirt, tronser, he holds bas, behind him there are three people sitting, in front of them there are some people standing. On the right there is a platform on that there is some people, buildings, plant. On the top left there is chinney, flowers, plants, wall and train. In the background there is railway track.







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.