Supplementary Material
Reasoning about Fine-grained Attribute Phrases using Reference Games

1. Annotation interface for user study

We gathered responses of human annotators for the task of the listener in the RG on Amazon mechanical turk using the interface shown in Fig. 1. Annotators are asked to select if the description refers to the "Left image", "Right image", or "I'm not sure". Each worker is paid $0.10 to annotate a single group consisting of 10 descriptions generated by speakers. Three workers are independently recruited for each task.

Figure 1. MTurk interface for user annotations

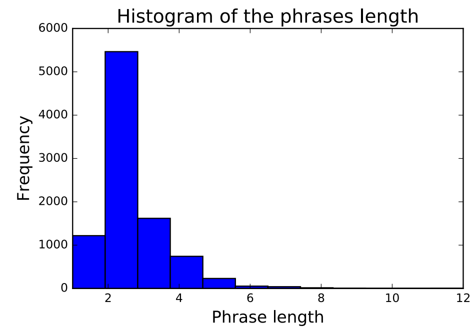

Figure 2. Histogram of the length of the attribute phrases in the training set of our dataset.

2. Additional dataset details

Some more details of the dataset are provided. Most of the attribute phrases have two words, and the longest is 12 words long. The histogram of the phrase lengths in the training set is shown in Figure 2. Additional examples of annotations are shown in Figure 3. Table 1 shows the top 20 most frequent attribute phrases, and attribute phrase pairs.

3. Additional results

Visualizing attribute phrases. Here we show more visualizations of the space of the attribute phrases. Figure 4 is the detailed version of Figure 6 in the paper. The embedded space of the contrastive phrases "P_1 vs. P_2" obtained using our discerning listener DL model in Figure 5. Figure 6 shows the embedding of images obtained by the SL. Phrases with similar semantic meanings, or images with same attributes, are clustered together.

Image retrieval with descriptive attributes. Figure 7 shows additional image retrieval results using SL (extension of Figure 7 in the paper.)

Comparing speakers Figure 8 for more examples that compare the simple and the pragmatic speakers. The last image pair is a challenging example where two images are very similar and the target image is misleading (the propellor looks like being on the wings but is in fact on the nose). SS fails on this case with most of generated phrases to be true to both images. DS successfully describes the major difference of wings and number of seats, and SL_r improves the ordering.

Attribute-based explanations for differences between two categories. Figure 9 shows additional examples of attributes generated as differences between two categories (more examples of Figure 8 in the paper.) The first and second example show that different phrases are generated for one category when it is compared to different categories. When compared with "A380", "Falcon 900" is considered small (DS generates "less windows"); When compared with "DR-400", "Falcon 900" is considered large (DS generates "large plane"). It reveals that DS has learnt the relative nature of phrases.

Figure 3. More example annotations from our dataset.

	Phrases	Freq.	Phrase pairs	Freq.
1	facing left	1258	facing right **VS** facing left	603
2	facing right	1214	facing left **VS** facing right	540
3	on the ground	785	on the ground **VS** in the air	198
4	private plane	647	in the air **VS** on the ground	165
5	small plane	550	commercial plane **VS** private plane	158
6	commercial plane	516	private plane **VS** commercial plane	155
7	in the air	458	large plane **VS** small plane	110
8	white color	402	on the ground **VS** flying in the air	104
9	white	376	propellor engine **VS** turbofan engine	98
10	turbofan engine	328	small plane **VS** big plane	92
11	propellor engine	310	big plane **VS** small plane	91
12	propeller engine	291	flying in the air **VS** on the ground	90
13	single engine	289	small **VS** large	87
14	on ground	288	in air **VS** on ground	85
15	flying in the air	281	outside **VS** inside	85
16	military plane	252	turbofan engine **VS** propellor engine	84
17	small	240	large **VS** small	83
18	large plane	238	small plane **VS** large plane	81
19	jet engine	233	on ground **VS** in air	81
20	big plane	233	inside **VS** outside	68

Table 1. Top 20 attribute phrases and contrastive attribute phrases from the training set in our dataset.

The last example is a challenging one with two very similar categories. The model fails in a pattern of describing undistinguishable attributes (engine, stabilizer) and attributes irrelative with categories (color, on ground or not). It also emphasizes that “757-200” is smaller than “A310” , but in fact they have similar size.

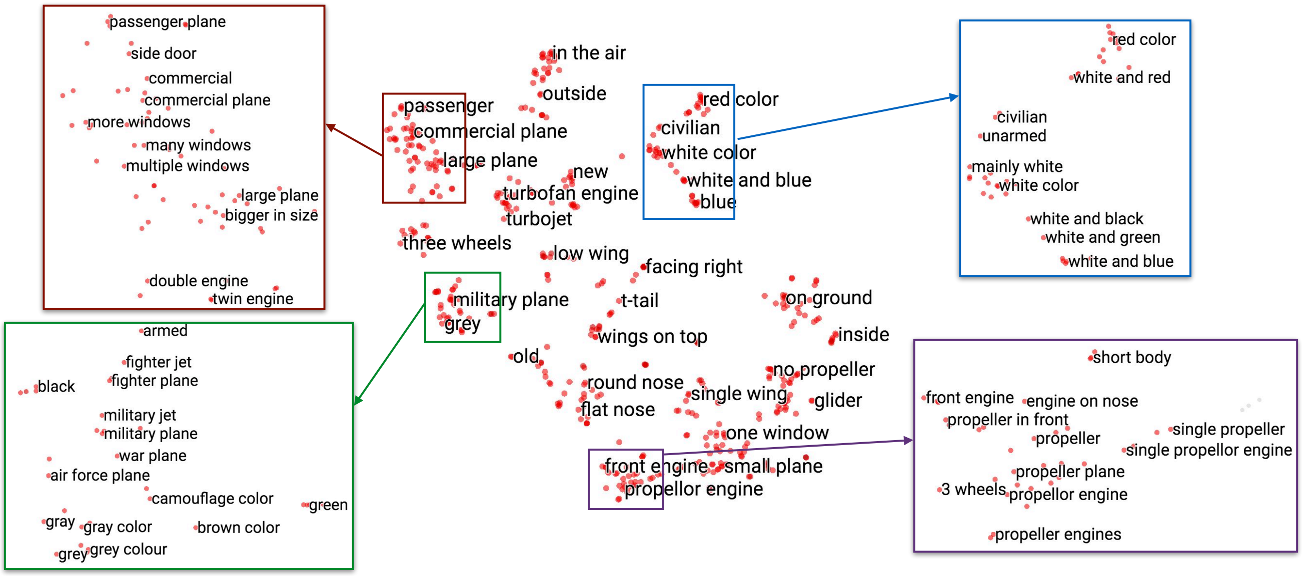

Figure 4. t-SNE embedding of attribute phrases from our simple listener (SL) model.

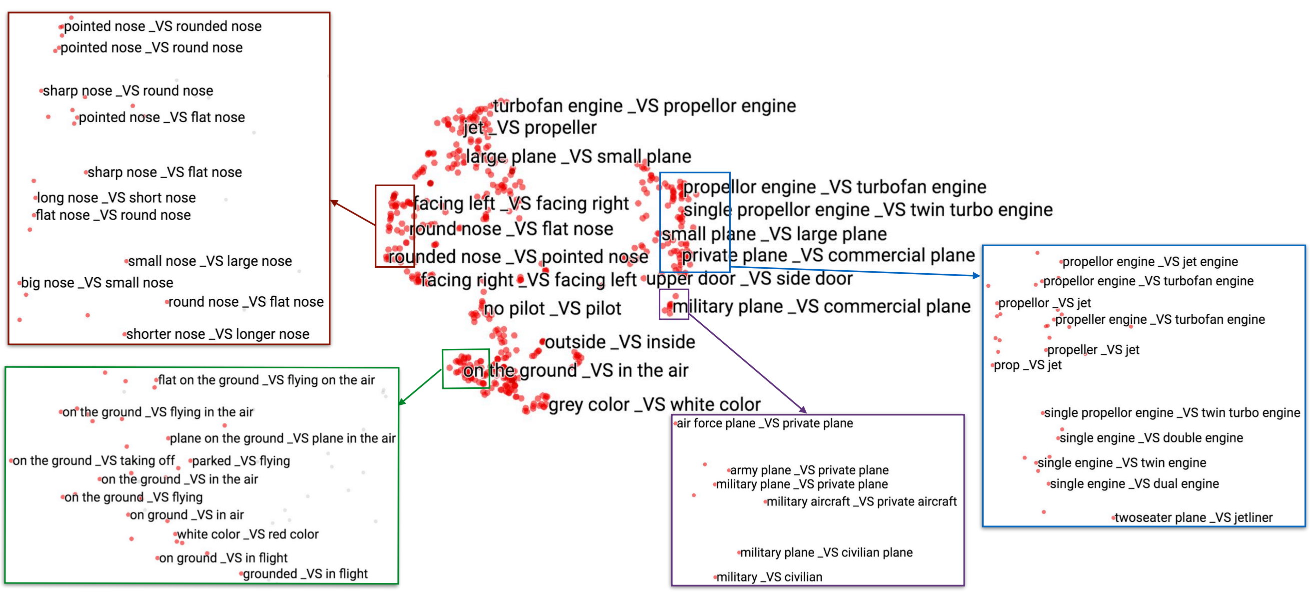

Figure 5. t-SNE embedding of contrastive attribute phrases, e.g. "P_1 vs. P_2", from our discerning listener (DL) model.

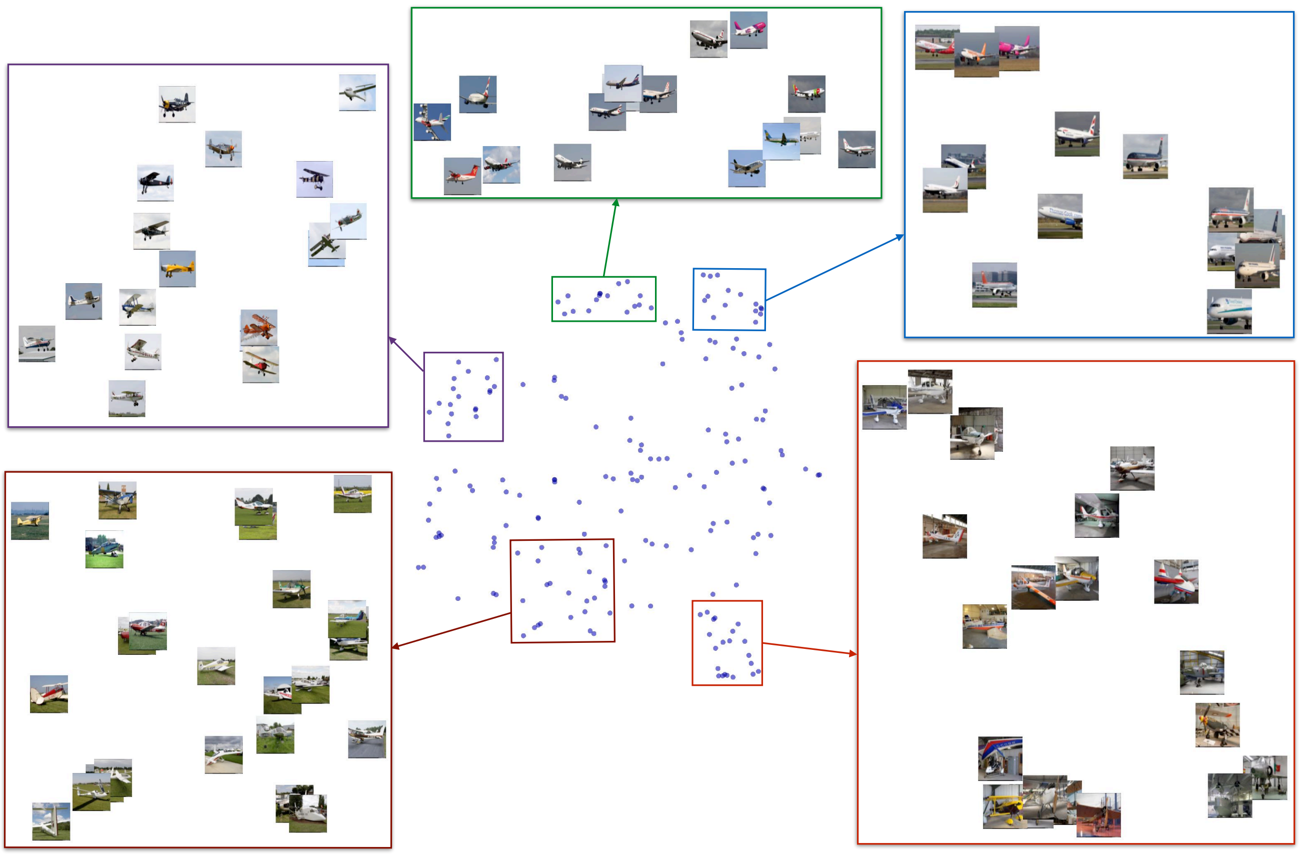

Figure 6. t-SNE embedding of 200 randomly selected images using our simple listener (SL) model. Images have same attributes are clustered together. For example, highlighted five boxes in the figure have the attribute “private plane on grass”, “private plane in the air”, “passenger plane in the air”, passenger plane on runway”, and “in hangar”.

Figure 7. Top 18 images ranked by the listener for various attribute phrases as queries (shown on top). We rank the images by the scores from the simple listener SL on the concatenation of the attribute phrases. The images are ordered from top to bottom, left to right.

SS:	SS + SL$_r$:	DS:	DS + SL$_r$:
? facing right	✔ fighter plane	? facing right	✔ single engine
✔ military plane	✔ small plane	✔ small plane	✔ two seater plane
✔ on ground	✔ military plane	✔ small	? propellor engine
✔ no windows on body	✔ no windows on body	✔ grey color	✔ small
✔ on the ground	✔ on the ground	✔ military plane	✔ small plane
✔ small plane	✔ on ground	✔ small plane	✔ small plane
? outside	? facing right	? propellor engine	✔ military plane
✔ grey	✔ grey	✔ single engine	✔ few windows
✔ jet engine	✔ jet engine	✔ few windows	? facing right
✔ fighter plane	? outside	✔ two seater plane	✔ grey color

SS:	SS + SL$_r$:	DS:	DS + SL$_r$:
✔ facing left	✔ jet fighter	✔ army plane	✔ jet engine
✔ jet engine	✔ fighter jet	✘ facing right	✔ on runway
✔ jet fighter	✔ jet engine	✔ military plane	✔ black in color
✔ military	✔ military plane	✔ jet engine	✔ military plane
✘ on the ground	✔ military	✔ gray color	✔ turbofan engine
✔ fighter jet	? turbofan engine	✔ facing left	✔ army plane
✔ military plane	✔ facing left	✔ on runway	✔ facing left
? turbofan engine	✘ on the ground	✔ turbofan engine	✔ gray in color
✘ single engine	✘ single engine	✔ gray in color	✔ gray color
✘ facing right	✘ facing right	✔ black in color	✘ facing right

SS:	SS + SL$_r$:	DS:	DS + SL$_r$:
? military plane	? military plane	? facing the left	? twin propellor
? propellor engine	? facing left	✘ three wheels	✔ 2 seater
? facing left	? open cockpit	? twin propellor	✔ 2 seater
? in flight	? military	✔ two seater	? facing the left
? open cockpit	? propellor engine	? green and yellow	✔ two seater
? _UNK between front wheels	? in flight	✔ one set of wings in front	✔ single wing
? single propellor engine	✘ single person aircraft	✔ 2 seater	✘ three wheels
✘ single person aircraft	? _UNK between front wheels	? no fan in nose	? no fan in nose
✔ no cone on nose	✔ no cone on nose	✔ single wing	✔ one set of wings in front
? military	? single propellor engine	✔ 2 seater	? green and yellow

Figure 8. More pragmatic speaker results. Given the image pairs in the left as input, we use SS and DS to generate phrases, and then use SL$_r$ to rerank them. SL$_r$ only takes the descriptions targeted at images in green boxes as input. Green checks mean human listener picks correct image with majority vote, X marks mean human listener picks opposite image with majority vote, and question marks mean human listener is uncertain which image is referred to.

Figure 9. More examples of attribute-based explanations for visual differences between two categories. Phrases are generated by DS and sorted by their occurrence frequency.