A. Impact of Pretrained Word Embeddings and Text Encoders

TransResNet encodes captions using a transformer architecture, which can be pre-trained:

- either by pre-training the word embeddings on a large corpus of text. In this case we used the pre-trained word vector released by FastText [5]
- or by pre-training the entire encoder on a similar task, in which case we followed the setting of [36].

Table 9, Table 10 and Table 12 show several ablation studies showing the importance of this pre-training.

The same word-pretraining can be attempted on generative models as well. Table 11 shows that 0.8 BLEU can be gained.

B. Engaging Captions, with no personality conditioning

Engaging-only Captions Instead of asking to author a caption based on a personality trait, we can ask humans to simply write an "engaging" caption instead, providing them with no personality cue. We found that human annotators overall preferred unconditioned captions to those conditioned on a personality by a slight margin ($\sim 54\%$). To further understand this difference, we split the images into three subsets based on the personality on which the PERSONALITY-CAPTIONS annotator conditioned their caption, i.e. whether the personality was positive, negative, or neutral. We then examined the engagingness rates of images for each of these subsets. In the set where PERSONALITY-CAPTIONS annotators were provided with positive personalities, which totaled 185 out of the 500 images, we found that human annotators preferred the captions conditioned on the personality to those that were not. However, in the other two sets, we found that the unconditioned captions were preferred to the negative or neutral ones. For these two subsets, we believe that, without the context of any personality, annotators may have preferred the inherently more positive caption provided by someone who was asked to be engaging but was not conditioned on a personality.

Diversity of captions We found that the captions written via our method were not only more engaging for positive personality traits, but also resulted in more diversity in terms of personality traits. To measure this diversity, we constructed a model that predicted the personality of a given comment. The classifier consists in the same Transformer as described in 4.3, pre-trained on the same large dialog corpus, followed by a softmax over 215 units. We then compare the total number of personality types as predicted by the classifier among each type of human-labeled data: "engaging" captions conditioned on personalities, "engaging" captions not conditioned on personalities, and traditional image captions. That is, we look at each caption given by the human annotators, assign it a personality via the classifier, and then look at the total set of personalities we have at the end for each set of human-labeled data. For example, out of the 500 human-generated traditional captions, the classifier found 63% of all possible positive personalities in this set of captions. As indicated in Table 14, the human annotators who were assigned a personality produce more diverse captions, particularly negatively and neutrally conditioned ones, as compared to human annotators who are just told to be "engaging" or those who are told to write an image caption.

C. Comparing Generative and Retrieval Models on COCO

The ultimate test of our generative and retrieval models on PERSONALITY-CAPTIONS is performed using human evaluations. Comparing them using automatic metrics is typically difficult because retrieval methods perform well with ranking metrics they are optimized for and generative models perform well with word overlap metrics they are optimized for, but neither of these necessarily correlate with human judgements, see e.g. [58].

Nevertheless, here we compare our generative and retrieval models directly with automatic metrics on COCO. We computed the BLEU, CIDEr, SPICE, and ROUGE-L scores for our best TransResNet model. The comparison is given in Table 15.

Model	Text Encoder Pretraining	Caption retrieval			
		R@1	R@5	R@10	Med Rank
		1k Images			
m-CNN [31]		42.8	-	84.1	2.0
UVS [25]		43.4	75.7	85.8	2.0
HM-LSTM [39]		43.9	-	87.8	2.0
Order Embeddings [49]		46.7	-	88.9	2.0
Embedding Net [51]		50.4	79.3	69.4	-
DSPE+Fisher Vector [52]		50.1	-	89.2	-
sm-LSTM [19]		53.2	83.1	91.5	1.0
VSE++ (ResNet, FT) [13]		64.6	90.0	95.7	1.0
GXN (i2t+t2i) [15]		68.5	-	**97.9**	1.0
[12]		**69.8**	**91.9**	96.6	1.0
Transformer†, Resnet152	Word	21.7	45.6	58.9	7.0
Bag of words, ResNeXt-IG-3.5B	None	51.6	85.3	93.4	1.4
Bag of words†, ResNeXt-IG-3.5B	Word	54.7	87.1	94.5	1.0
Transformer, ResNeXt-IG-3.5B	None	63.4	90.6	96.3	1.0
Transformer†, ResNeXt-IG-3.5B	Word	66.6	90.6	96.3	1.0
Transformer*, ResNeXt-IG-3.5B	Full	67.3	91.7	96.5	1.0
		1k Images			
Order Embeddings [49]		23.3	-	65.0	5.0
VSE++ (ResNet, FT) [13]		41.3	71.1	81.2	2.0
GXN (i2t+t2i) [15]		42.0	-	84.7	2.0
Transformer, Resnet152	Word	7.8	21.9	31.2	30.0
Bag of words, ResNeXt-IG-3.5B	None	26.6	58.6	73.0	4.0
Bag of words, ResNeXt-IG-3.5B	Word	29.7	62.9	75.7	3.0
Transformer, ResNeXt-IG-3.5B	None	38.8	71.6	82.7	2.0
Transformer, ResNeXt-IG-3.5B	Word	44	73.7	**84**	2.0
Transformer, ResNeXt-IG-3.5B	Full	**44.3**	**74.5**	83.9	2.0

Table 9: More detailed results for retrieval model performance on COCO Captions using the splits of [24]. For our TransResNet models, we compare two types of pretraining: Full indicates a model with a pretrained text encoder, while Word indicates a model with pretrained word embeddings only.

Model	Text Encoder Pretraining	Caption retrieval			
		R@1	R@5	R@10	Med Rank
UVS [25]		23.0	50.7	62.9	5.0
UVS (Github)		29.8	58.4	70.5	4.0
Embedding Net [51]		40.7	69.7	79.2	-
DAN [38]		41.4	73.5	82.5	2.0
sm-LSTM [19]		42.5	71.9	81.5	2.0
2WayNet [11]		49.8	67.5	-	-
VSE++ (ResNet, FT) [13]		52.9	80.5	87.2	1.0
DAN (ResNet) [38]		55.0	81.8	89.0	1.0
GXN (i2t+t2i) [15]		56.8	-	89.6	1.0
Transformer, Resnet152	Word	10.3	27.3	38.8	19
Bag of words, ResNeXt-IG-3.5B	None	50.0	81.1	90.0	1.5
Transformer, ResNeXt-IG-3.5B	None	55.6	83.2	90.5	1.0
Bag of words, ResNeXt-IG-3.5B	Word	58.6	87.2	92.9	1.0
Transformer, ResNeXt-IG-3.5B	Full	62.3	88.5	94.4	1.0
Transformer, ResNeXt-IG-3.5B	Word	**68.4**	**90.6**	**95.3**	1.0

Table 10: Retrieval model performance on Flickr30k using the splits of [24]. For our models, we compare two types of pretraining: Full indicates a model with a pretrained text encoder, while Word indicates a model with pretrained word embeddings only.

Method	Image Encoder	Personality	BLEU1	BLEU4	ROUGE-L	CIDEr	SPICE
no pretraining:							
SHOWTELL	ResNeXt-IG-3.5B	Yes	38.4	7.3	24.3	9.6	1.6
SHOWATTTELL	ResNeXt-IG-3.5B	Yes	43.3	7.1	27.0	12.6	3.6
UPDOWN	ResNeXt-IG-3.5B	Yes	44.0	8.0	27.4	**16.5**	**5.2**
with word embedding pretraining:							
SHOWTELL †	ResNeXt-IG-3.5B	Yes	40.1	7.7	25.3	11.0	2.2
SHOWATTTELL †	ResNeXt-IG-3.5B	Yes	44.6	7.5	25.9	12.6	3.6
UPDOWN †	ResNeXt-IG-3.5B	Yes	**44.8**	**8.1**	**27.7**	16.3	**5.2**

Table 11: Comparing Generative model caption performance on the PERSONALITY-CAPTIONS test set: pretrained word embeddings vs. no pretraining. Pretraining makes a very small impact in this case, unlike in our retrieval models.

Text Encoder				
Encoder Type	Pretraining	Image Encoder	Personality Encoder	R@1
Transformer	Full	ResNeXt-IG-3.5B	Yes	**77.5**
Transformer	Word	ResNeXt-IG-3.5B	Yes	71.7
Bag of Words	Word	ResNeXt-IG-3.5B	Yes	66.2
Transformer	None	ResNeXt-IG-3.5B	Yes	65.9
Bag of Words	None	ResNeXt-IG-3.5B	Yes	58.6
Transformer	Full	ResNeXt-IG-3.5B	No	53.9
Transformer	Full	Resnet152	Yes	51.7
Transformer	Word	Resnet152	Yes	45.4
Transformer	None	Resnet152	Yes	40.6
Bag of Words	Word	Resnet152	Yes	40.5
Bag of Words	None	Resnet152	Yes	35.4
Transformer	Full	Resnet152	No	18.7

Table 12: Retrieval model performance on PERSONALITY-CAPTIONS. We compare two types of pretraining: Full indicates a model with a pretrained text encoder, while Word indicates a model with pretrained word embeddings only.

Type of caption A	WIN PERCENTAGE		Type of caption B
Human (all) personality captions	45.5	**54.5**	Human engaging captions
Human (positive) personality captions	**51.2**	48.8	Human engaging captions

Table 13: Pairwise win rates of various approaches, evaluated in terms of engagingness

Annotation Task	**Personality Trait Coverage**		
	Positive	Neutral	Negative
Given Personalities	100%	100%	99.0%
Traditional Caption	63.0%	83.3%	47.0%
Engaging, No Conditioning	81.5%	91.7%	71.4%
PERSONALITY-CAPTIONS	82.7%	94.4%	87.8%

Table 14: Caption diversity in human annotation tasks. PERSONALITY-CAPTIONS provides more diverse personality traits than traditional captions or collecting engaging captions without specifying a personality trait to the annotator, as measured by a personality trait classifier.

Model	BLEU1	BLEU4	ROUGE-L	CIDEr	SPICE
TransResNet	50.6	10.9	38.0	49.1	13.9
SHOWTELL	78.2	35.0	56.6	119.9	20.8
SHOWATTTELL	78.8	35.6	57.1	121.8	20.6
UPDOWN	79.3	**36.4**	**57.5**	**124.0**	21.2

Table 15: Generative and retrieval model performance on COCO caption using the test split of [24]. All models use ResNeXt-IG-3.5B image features.

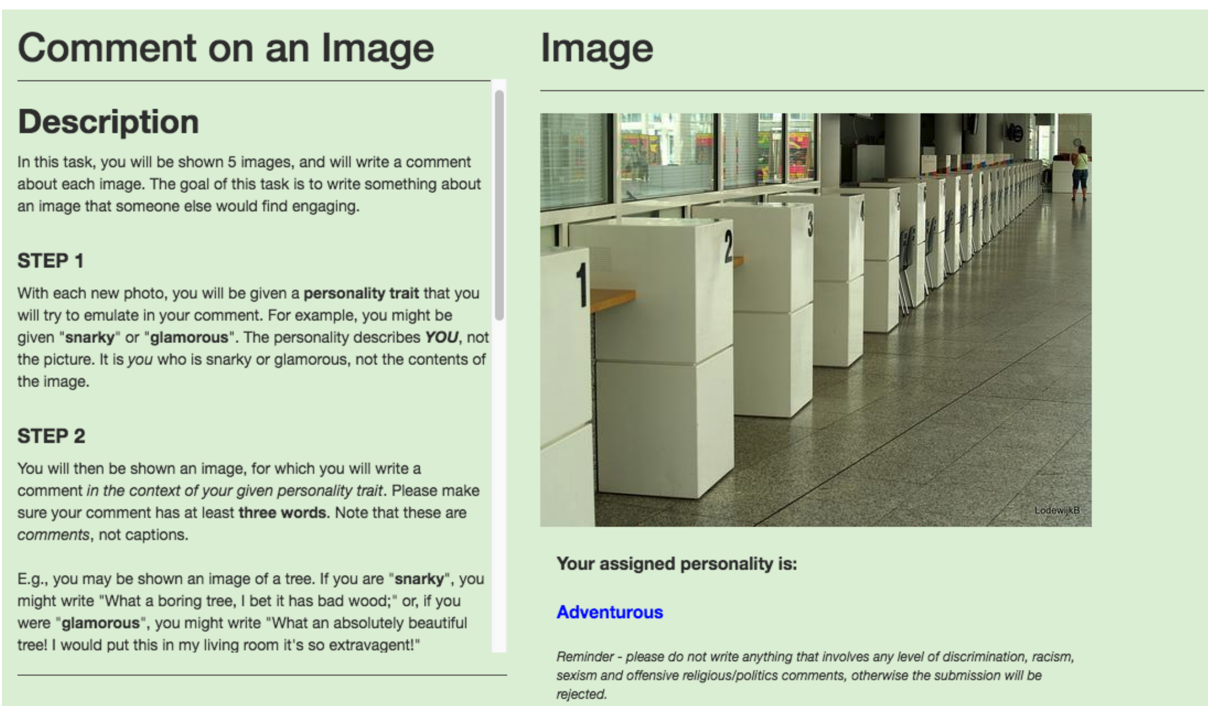


Figure 3: Instructions for the annotation task collecting the data for PERSONALITY-CAPTIONS.

Sarcastic
Yes please sit by me

Mellow
Look at that smooth easy catch of the ball. like ballet.

Zany
I wish I could just run down this shore!

Contradictory
Love what you did with the place!

Mellow
Look at that smooth easy catch of the ball. like ballet.

Energetic
About to play the best tune you've ever heard in your life. Get ready!

Kind
they left me a parking spot

Spirited
That is one motor cycle enthusiast!!!

Creative
Falck alarm, everyone. Just a Falck alarm.

Crazy
I drove down this road backwards at 90 miles per hour three times

Morbid
I hope this car doesn't get into a wreck.

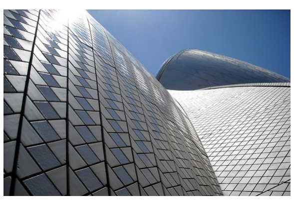
Questioning
Why do people think its cool to smoke cigarettes?

Table 16: Some samples from PERSONALITY-CAPTIONS. For each sample we asked a person to write a caption that fits both the image and the personality.

Old-fashioned
origin: TransResNet
fit: does not fit image
Each of these hammers has a mission.

Destructive
origin: TransResNet
fit: does not fit personality
that dog is going to drown! someone save it.

Courageous
origin: TransResNet
fit:neither
Look at all of those sewing materials! You could create all sorts of art projects with them!



Meticulous
origin: human
fit: neither
The desert is so overwhelming and vast I totally want to go exploring again!

Sympathetic
origin: human
fit: does not fit personality
relaxing,calm and authentic

Bewildered
origin: human
fit:neither
Graduating school and you finally feel like you're invincible.

Table 17: Some examples of captions that do not fit either the personality or the image, produced by humans and TransResNet

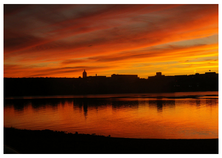
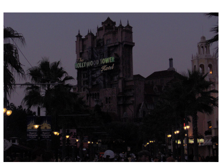
Image and Pers.	Use pers.	Captioning	Caption
	No	Standard	A city on the background, a lake on the front, during a sunset.
	No	Engaging	Talk about summer fun! Can I join? :)
	Yes	Human	i feel moved by the sunset
	Yes	TransResNet	The water at night is a beautiful sight.
Spirited	Yes	UPDOWN	This is a beautiful sunset!
	No	Standard	Rose colored soft yarn.
	No	Engaging	I really want to untangle that yarn.
	Yes	Human	I cannot believe how yummy that looks.
	Yes	TransResNet	What is up with all the knitting on my feed
Ridiculous	Yes	UPDOWN	I would love to be a of that fruit!
	No	Standard	A beautiful mesa town built into the cliffs.
	No	Engaging	That is a strange cave
	Yes	Human	It must be very dangerous if children play there
	Yes	TransResNet	I hope my kids don't climb on this.
Maternal	Yes	UPDOWN	I hope this is a beautiful place.
	No	Standard	Hockey players competing for control of the hockey puck.
	No	Engaging	Great save, goalie!!
	Yes	Human	Hockey is a little too barbaric for my taste.
	Yes	TransResNet	Hockey players gracefully skate across the ice.
Sophisticated	Yes	UPDOWN	This hockey is like they are a great of the game.
	No	Standard	Hollywood Tower at Night
	No	Engaging	I went to that theme park, but was too scared to get on that ride!
	Yes	Human	I am so excited to be here!
	Yes	TransResNet	I remember going to disney world, it was one of the best trips I've ever done.
Happy	Yes	UPDOWN	This looks like a beautiful view!

Table 18: Example variants of the captions shown to human annotators in the human evaluation tasks in Section 5.3. The first two captions are human annotations not conditioned on a personality; the next three are captions conditioned on the listed personality, and are generated via a human annotator, TransResNet, and UPDOWN respectively.

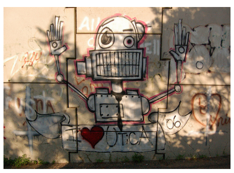
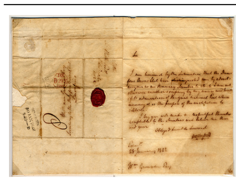
Image	Personality	Generated comment
	Sweet	What a cute puppy, reminds me of my friends.
	Skeptical	I don’t think this dog will bite me.
	Sympathetic	poor dog! It looks so hungry :c
	Vague	it’s a dog
	Wishful	I wish that I had a dog as cute as him.
	Cultured	I love a cultural celebration.
	Skeptical	I’m not sure if these are guys in costumes or time travelers.
	Sweet	I love that they are celebrating their traditions and culture.
	Overimaginative	They look like they could be dancers in a fantasy movie with dragons!
	Sympathetic	I feel sorry for him having to wear that
	Romantic	If I was an insect, I would definitely make this my mate.
	Humble	I am grateful that spiders eat these disgusting bugs.
	Paranoid	What is going on? Are these insects dangerous?
	Creative	I made something like this from colored toothpicks once
	Money-minded	how much are those? those looks expensive
	Happy	That is so cool! I I love street art!
	Optimistic	The future is bright for people who can dream in artistic ways.
	Critical	I do believe this taggers verbage is a tad junvenile
	Charming	What a charming wall.
	Adventurous	I think I could create art like that, I will go learn and take action.
	Adventurous	I am so ready for the conference.
	Cultured	This conference is one of the most important ones in the country.
	Vague	The organization on that table is uncertain.
	Dramatic	OMG!! This ceremony is frightening!
	Sympathetic	I feel bad for these people being so cramped in this room.
	Old-fashioned	Such old fashioned script, a true lost art.
	Charming	I could use these to write to my loved ones.
	Argumentative	Can you even read this through all the jpeg artifacts?
	Anxious	I hope this paper doesnt tear, history will be destroyed.
	Dramatic	Some of the most profound things ever written have been on linen.
	Wishful	I wish I could have a life as easy as a plant.
	Money-minded	This plant is probably worth a lot of money
	Critical	the leaf is ruining the picture
	Humble	This plant is a symbol of life in humble opinion. Just gorgeous!
	Paranoid	If you eat this leaf it definetly will not poison you. Or will it...
	Romantic	This valentine concert is for lovers.
	Boyish	It’s always fun to get down and jam with the boys!
	Creative	musician performing a song of theirs
	Sweet	oh what lovely young musicians
	Money-minded	I wonder how much the musicians have in student loan debt.

Table 19: More example predictions from our best TransResNet model on the Personality-Captions validation set.