A Accuracy per Question Types

For further analysis, we report the performance of the model for each question type. All questions are categorized based on the first two words, and the top 10 frequent question types are provided as per this categorization. The results are shown in Table A.1.

For most question types, our method shows higher accuracy and lower mean rank than the methods used for comparison. The performance of the statistical methods drops to almost zero for some question types. For example, the performance of the statistical methods for some question types where the answer is a common word, such as "Is the" or "Are there," is very poor.

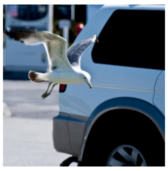
B Qualitative Results

Additional examples of qualitative results from applying the proposed model and comparison methods to the GQA and FSVQA datasets are shown in Figure B.1.

GQA		Ours		TF-IDF		YAKE		EmbedRank	
	num.	Acc.	Mean Rank	Acc.	Mean Rank	Acc.	Mean Rank	Acc.	Mean Rank
Is the	22,507	**0.437**	**1.898**	0.058	3.819	0.326	–	0.104	2.880
What is	20,435	0.544	1.444	**0.581**	**1.295**	0.237	–	0.495	1.353
Are there	13,004	0.276	3.217	0.000	4.859	0.000	–	**0.395**	**2.020**
Who is	6,452	0.366	**1.773**	0.184	2.158	**0.940**	–	0.080	2.541
Is there	5,270	0.269	2.971	0.000	4.792	0.000	–	**0.314**	**2.272**
Does the	5,236	**0.393**	**2.108**	0.031	3.812	0.002	–	0.126	2.502
On which	5,121	0.369	**1.701**	0.001	2.074	**0.997**	–	0.045	3.060
Do you	4,976	0.291	2.936	0.005	4.770	0.000	–	**0.338**	**2.118**
Which kind	4,410	**0.744**	**1.129**	0.720	1.143	0.033	–	0.471	1.370
What kind	4,148	**0.719**	**1.171**	0.704	1.175	0.060	–	0.476	1.358

FSVQA		Ours		TF-IDF		YAKE		EmbedRank	
	num.	Acc.	Mean Rank	Acc.	Mean Rank	Acc.	Mean Rank	Acc.	Mean Rank
How many	11,592	**0.271**	**2.579**	0.075	2.926	0.003	–	0.004	3.930
What is	10,394	0.343	2.564	0.542	1.724	0.099	–	**0.551**	**1.698**
What color	10,237	**0.678**	**1.393**	0.073	2.385	0.097	–	0.110	2.133
Is the	5,596	**0.336**	**2.312**	0.087	4.788	0.046	–	0.163	2.452
Is this	4,475	**0.346**	2.172	0.075	5.662	0.074	–	0.279	**2.102**
What are	2,045	0.273	2.568	**0.603**	**1.577**	0.007	–	0.602	1.587
What kind	1,695	0.245	2.393	**0.889**	**1.160**	0.588	–	0.810	1.232
Are the	1,473	**0.291**	**2.367**	0.075	5.720	0.025	–	0.131	2.499
What type	1,113	0.232	2.431	**0.864**	**1.190**	0.543	–	0.812	1.226
Where is	1,106	0.500	2.042	**0.568**	**1.626**	0.062	–	0.397	1.814

Table A.1: Performance per question types. The upper and lower tables show the results of GQA and FSVQA, respectively.

GQA				
Question	In which part of the image is the car, the bottom or the top?	What is on the chair the backpack is to the left of?	Where is that snowboard?	What is this bird in?
Full-sentence Answer	The car is in the top of the image.	The jacket is on the chair.	The snowboard is on the snow.	The bird is in the air.
GT Keyword	top	jacket	snow	air
TF-IDF	**top**	**jacket**	snowboard	**air**
YAKE	car	**jacket**	snowboard	bird
EmbedRank	car	**jacket**	snowboard	bird
Ours	**top**	**jacket**	snowboard	bird

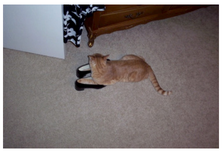
FSVQA				
Question	Is this cat being aggressive?	What is the flooring made of?	How many sections of bridge can you see?	What is the man trying to catch?
Full-sentence Answer	No, this cat is not being aggressive.	The flooring is made of tile.	3 sections of bridge i can see.	The man is trying to catch frisbee.
GT Keyword	no	tile	3	frisbee
TF-IDF	aggressive	flooring	sections	catch
YAKE	this	flooring	sections	man
EmbedRank	cat	flooring	bridge	**frisbee**
Ours	**no**	**tile**	bridge	catch

Figure B.1: Examples of keyword extraction results using the GQA and FSVQA datasets. The upper and lower tables show the results of GQA and FSVQA, respectively.