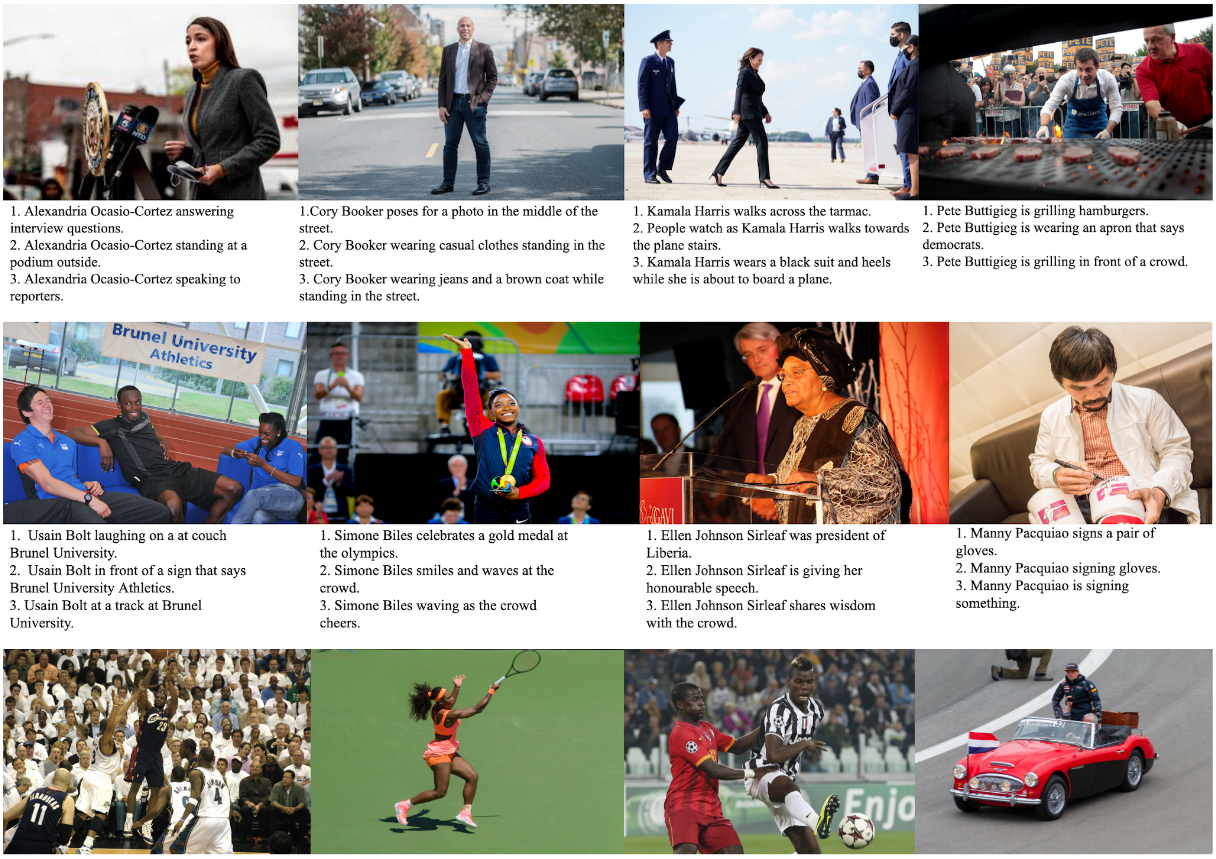
Towards Multimodal Vision-Language Models Generating Non-Generic Text — Supplementary Material

Wes Robbins
Montana State University
wesley.robbins10@gmail.com

Zanyar Zohourianshahzadi & Jugal Kalita
University of Colorado, Colorado Springs
{zzohouri,jkalita}@uccs.edu

Figure 1: More samples from the Politicians and Athletes Dataset. Well-known individuals can be seen in a variety of scenes.



1. Alexandria Ocasio-Cortez answering interview questions.
2. Alexandria Ocasio-Cortez standing at a podium outside.
3. Alexandria Ocasio-Cortez speaking to reporters.

1.Cory Booker poses for a photo in the middle of the street.
2. Cory Booker wearing casual clothes standing in the street.
3. Cory Booker wearing jeans and a brown coat while standing in the street.

1. Kamala Harris walks across the tarmac.
2. People watch as Kamala Harris walks towards the plane stairs.
3. Kamala Harris wears a black suit and heels while she is about to board a plane.

1. Pete Buttigieg is grilling hamburgers.
2. Pete Buttigieg is wearing an apron that says democrats.
3. Pete Buttigieg is grilling in front of a crowd.

1. Usain Bolt laughing on a at couch Brunel University.
2. Usain Bolt in front of a sign that says Brunel University Athletics.
3. Usain Bolt at a track at Brunel University.

1. Simone Biles celebrates a gold medal at the olympics.
2. Simone Biles smiles and waves at the crowd.
3. Simone Biles waving as the crowd cheers.

1. Ellen Johnson Sirleaf was president of Liberia.
2. Ellen Johnson Sirleaf is giving her honourable speech.
3. Ellen Johnson Sirleaf shares wisdom with the crowd.

1. Manny Pacquiao signs a pair of gloves.
2. Manny Pacquiao signing gloves.
3. Manny Pacquiao is signing something.

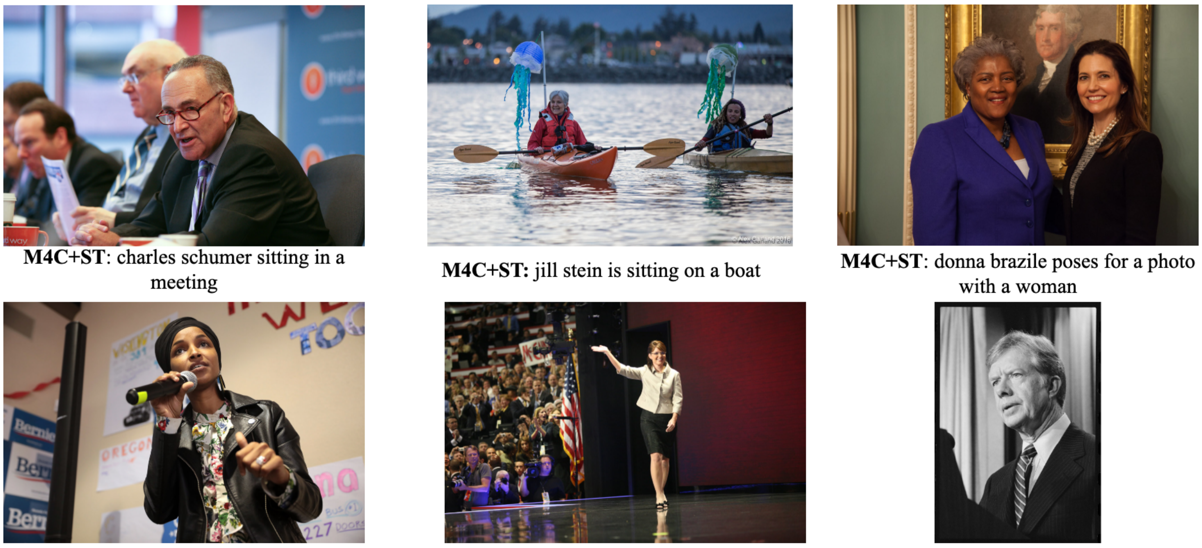
1. Lebron James throwing a basketball while wearing a dark Cleveland jersey.
2. Lebron James throwing the basketball and wearing a Cleveland jersey.
3. Lebron James throwing the basketball while wearing a Cleveland Jersey.

1. Serena Williams playing tennis on a tennis court in an orange outfit.
2. Serena Williams wearing an orange outfit while holding the tennis racket in her hand.
3. Serena Williams playing tennis while wearing an orange outfit.

1. Another player touches Paul Pogba as he kicks the ball.
2. Paul Pogba has his cleats up as he kicks the ball.
3. Paul Pogba wears a black and white uniform and yellow cleats.

1. Max Verstappen sitting in a car the has a French flag.
2. Max Verstappen sitting in the back of an old orange car.
3. Max Verstappen riding around the track before a race.

Figure 2: The *The Special Approach* allows zero-shot adaptation for individuals never seen in training by the captioning model. All individuals in the below images are not present in the PAC training set, yet our model (M4C+ST) is able to integrate the names into the captions at inference.



M4C+ST: charles schumer sitting in a meeting

M4C+ST: jill stein is sitting on a boat

M4C+ST: donna brazile poses for a photo with a woman

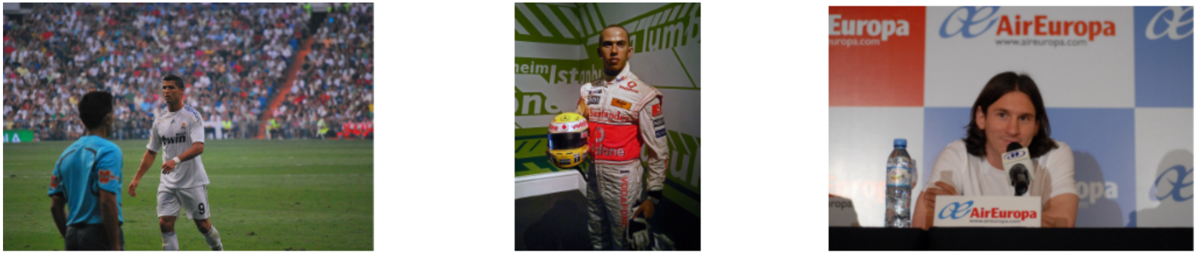
M4C+ST: ilhan omar holding a microphone

M4C+ST: sarah palin on a stage with a crowd

M4C+ST: a black and white photo of jimmy carter

Figure 3: Captions on PAC images. The special token model offers richer captions by utilizing person names. The vanilla M4C model and weights used below come from the MMF repository (Singh et al., 2020)

.



M4C+ST: cristiano ronaldo is on the field
M4C: a man in a baseball uniform with the number 5 on his back

M4C+ST: lewis hamilton is wearing a jersey that says vodaphone
M4C: a man wearing a jersey that says 'vodais' on it

M4C+ST: lionel messi standing infront of a sign that says aireuropa
M4C: a woman is wearing is a sign that says 'aireuropa'

Table 1: M4C+ST model performance on both PAC and TextCaps (extension of table in main paper). We find that by training on both datasets, we can get good performance of both PAC which focuses on person names and TextCaps which focuses on OCR tokens. In the training column, a → between datasets indicates one dataset was trained on before the other where as a comma in between datasets indicates they were trained on simultaneously. Datasets trained on simultaneously are followed by a sampling ratio between datasets(e.g PAC,TextCaps[1:8] is 1 batch of PAC per 8 batches of TextCaps).

			Metrics				
#	Training	Test	B-4	M	R	C	S
1 a.	TextCaps	PAC	0.7	5.1	11.7	10.4	3.0
b.		TextCaps	22.9	22.1	46.0	89.7	15.3
2 a.	PAC	PAC	8.6	13.3	29.1	98.9	17.6
b.		TextCaps	2.0	7.5	22.0	8.5	2.6
3 a.	TextCaps→PAC	PAC	9.1	14.8	30.4	102.6	18.7
b.		TextCaps	20.7	20.1	43.0	80.4	13.4
4 a.	PAC,TextCaps[1:8]	PAC	8.4	14.5	30.3	103.7	17.5
b.		TextCaps	22.1	20.9	45.3	84.5	24.0
5 a.	TextCaps→PAC,TextCaps[1:1]	PAC	5.1	12.8	25.7	73.0	14.8
b.		TextCaps	23.2	22.0	46.2	91.0	15.1

B-4: BLEU-4; M: METEOR; R: ROUGUE; C: CIDEr, S: SPICE

References

Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2020. Mmf: A multimodal framework for vision and language research. https://github.com/facebookresearch/mmf.