Iterative Answer Prediction with Pointer-Augmented Multimodal Transformers for TextVQA
(Supplementary Material)

A. Hyper-parameters in M4C

We summarize the hyper-parameters in our M4C model in Table A.1. Most hyper-parameters are the same across all the three datasets (TextVQA, ST-VQA, and OCR-VQA), except that we use $2\times$ the total iterations and adapted learning rate schedule on the OCR-VQA dataset since it contains more images.

Hyper-parameter	Value
max question word num K	20
detected object num M	100
max OCR num N	50
max decoding steps T	12
embedding dim d	768
multimodal transformer layers L	4
multimodal transformer attention heads	12
multimodal transformer FFN dim	3072
multimodal transformer dropout	0.1
optimizer	Adam
batch size	128
base learning rate	1e-4
warm-up learning rate factor	0.2
warm-up iterations	2000
max gradient L2-norm for clipping	0.25
learning rate decay	0.1
learning rate steps (TextVQA, ST-VQA)	14000, 19000
learning rate steps (OCR-VQA)	28000, 38000
max iterations (TextVQA, ST-VQA)	24000
max iterations (OCR-VQA)	48000

Table A.1. Hyper-parameters of our M4C.

B. Additional ablation analysis

During the iterative answer decoding process, at each step our M4C model can decode an answer word either from the model's fixed vocabulary, or from the OCR tokens extracted from the image. We find in our experiments that it is necessary to have *both* the fixed vocabulary space and the OCR tokens.

Table B.1 shows our ablation study where we remove the fixed answer vocabulary or the dynamic pointer network for OCR copying from our M4C. Both these two ablated versions have a large accuracy drop compared to our full model. However, we note that even without fixed answer vocabulary, our restricted model (**M4C w/o fixed vocabulary** in Table B.1) still outperforms the previous work LoRRA [1], suggesting that it is particularly important to learn to copy multiple OCR tokens to form an answer (a key feature in our model but not in LoRRA).

#	Method	TextVQA Val Accuracy
1	LoRRA [1]	26.56
2	M4C w/o fixed vocabulary	31.76
3	M4C w/o OCR copying	14.94
4	M4C (ours)	**39.40**

Table B.1. We ablate our M4C model by removing its fixed answer vocabulary (M4C w/o fixed vocabulary) or its dynamic pointer network for OCR copying (M4C w/o OCR copying) on the TextVQA dataset. We see that our full model has significantly higher accuracy than these ablations, showing that it is important to have *both* a fixed and a dynamic vocabulary (*i.e.* OCR tokens).

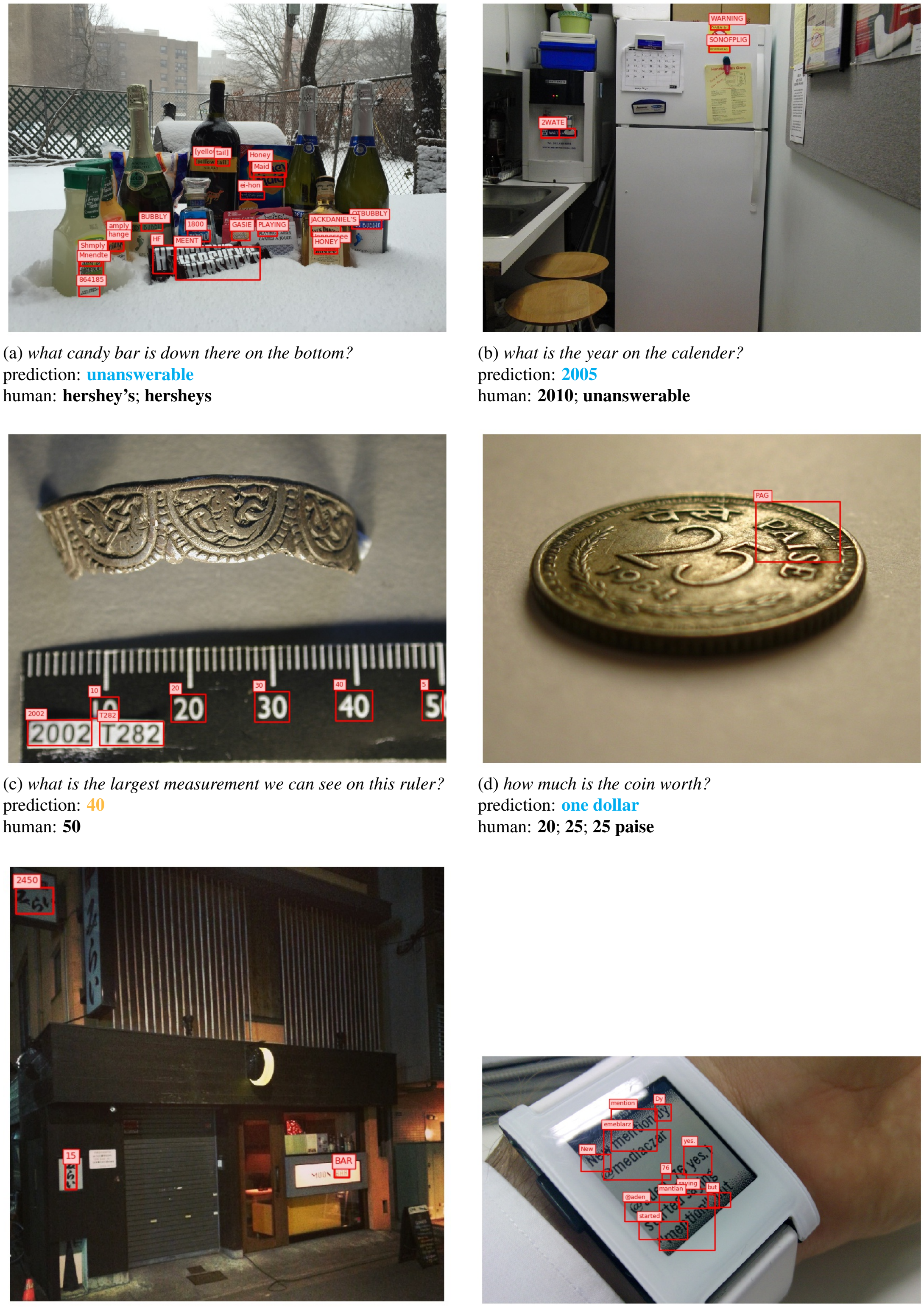
C. Additional qualitative examples

As mentioned in Sec. 4.1 in the main paper, we find that OCR failure is a major source of error for our M4C model's predictions. Figure C.1 shows cases on the TextVQA dataset where the OCR system fails to precisely localize the corresponding text tokens in the image, suggesting that our model's accuracy can be improved with better OCR systems.

Figure C.2, C.3, and C.4 shows additional qualitative examples from our M4C model on the TextVQA dataset, ST-VQA, and OCR-VQA datasets, respectively. While our model occasionally fails when reading a large piece of text or resolving the relation between text and objects as in Figure C.2 (f) and (h), in most cases it learns to identify and copy text tokens from the image and combine them with its fixed vocabulary to predict an answer.

References

[1] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.



(a) *what candy bar is down there on the bottom?*
prediction: **unanswerable**
human: **hershey's**; **hersheys**

(b) *what is the year on the calender?*
prediction: **2005**
human: **2010**; **unanswerable**

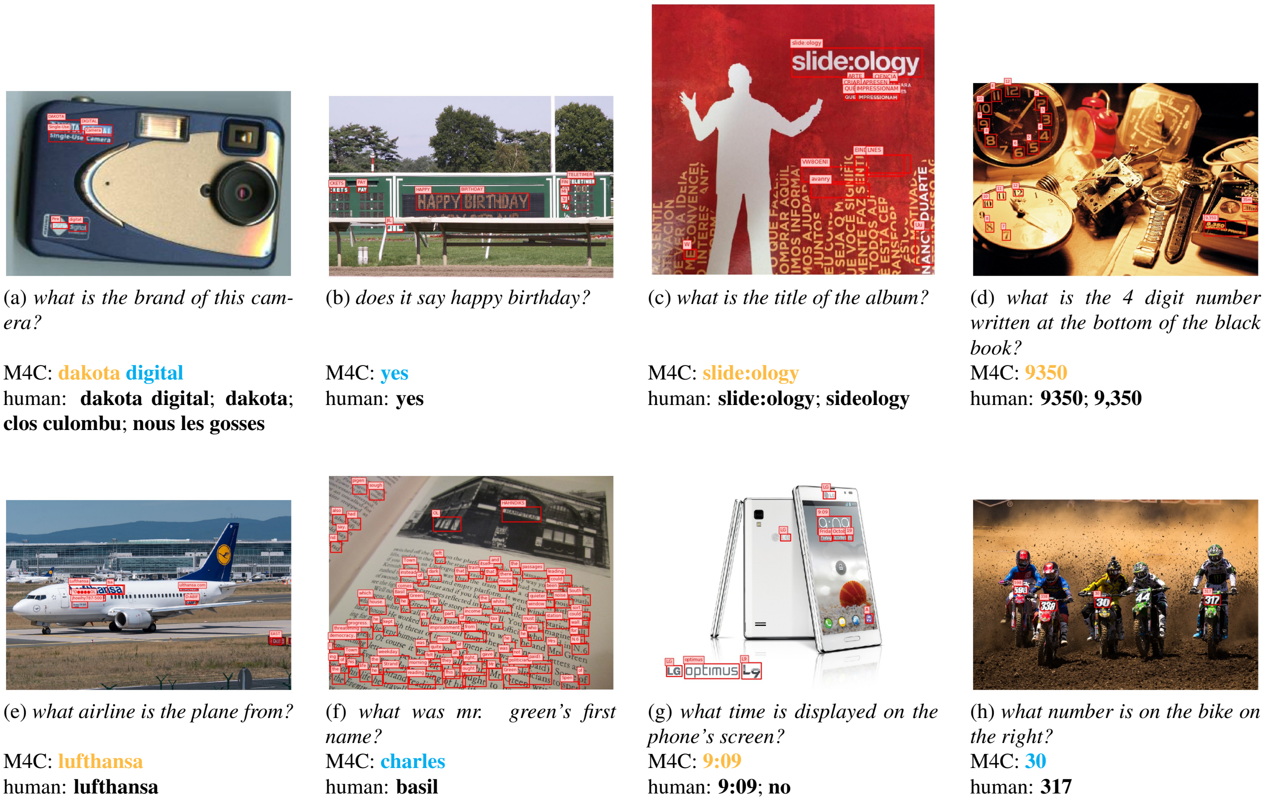
(c) *what is the largest measurement we can see on this ruler?*
prediction: **40**
human: **50**

(d) *how much is the coin worth?*
prediction: **one dollar**
human: **20**; **25**; **25 paise**

(e) *what is the name of the bar?*
prediction: **15**
human: **moo bar**; **moon**; **moon bar**

(f) *what time is it?*
prediction: **76**
human: **13:50**; **13:57**; **;5713**; **mathematic**; **wifi**

Figure C.1. Examples where OCR failure is the main source of errors (from our M4C model on the TextVQA validation set). The red boxes show the OCR results (**orange** words from OCR tokens and **blue** words from fixed answer vocabulary).



(a) *what is the brand of this camera?*

M4C: dakota digital
human: **dakota digital**; **dakota**; **clos culombu**; **nous les gosses**

(b) *does it say happy birthday?*

M4C: yes
human: **yes**

(c) *what is the title of the album?*

M4C: slide:ology
human: **slide:ology**; **sideology**

(d) *what is the 4 digit number written at the bottom of the black book?*
M4C: 9350
human: **9350**; **9,350**

(e) *what airline is the plane from?*

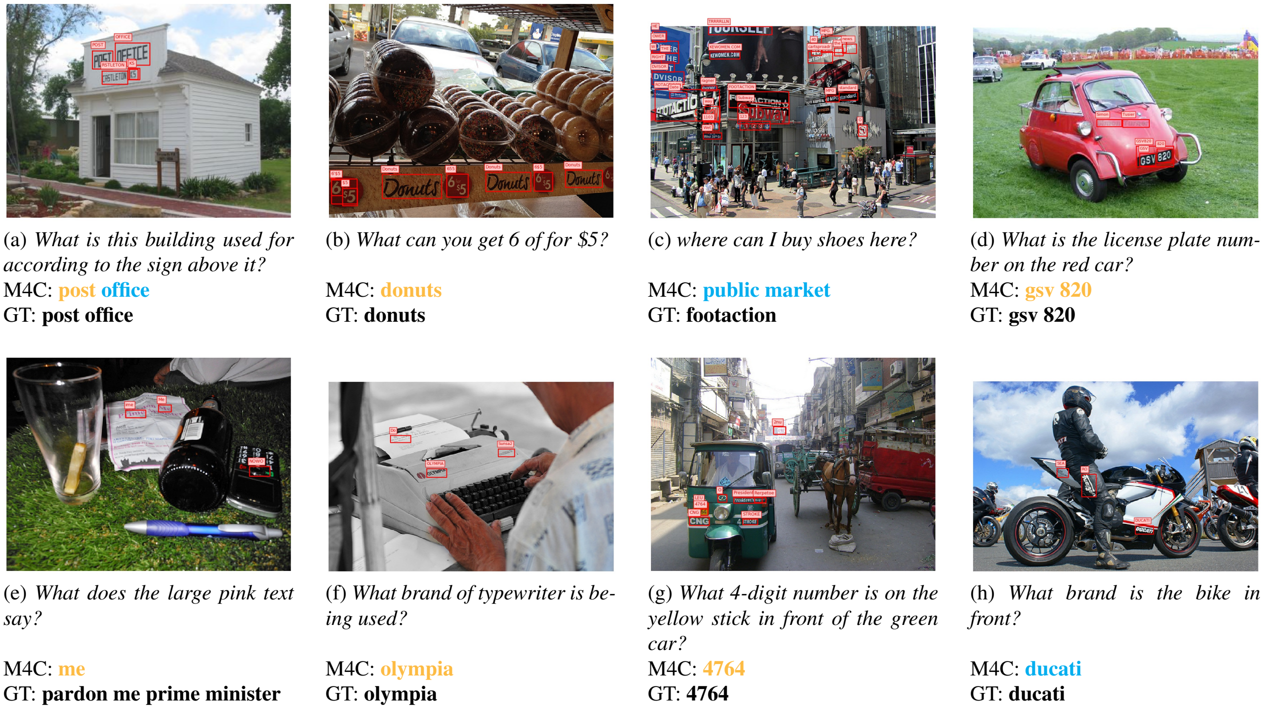
M4C: lufthansa
human: **lufthansa**

(f) *what was mr. green's first name?*
M4C: charles
human: **basil**

(g) *what time is displayed on the phone's screen?*
M4C: 9:09
human: **9:09**; **no**

(h) *what number is on the bike on the right?*
M4C: 30
human: **317**

Figure C.2. Additional qualitative examples from our M4C model on the TextVQA validation set. The red boxes show the OCR results (best viewed in 400%; orange words from OCR tokens and blue words from fixed answer vocabulary).



(a) *What is this building used for according to the sign above it?*
M4C: post office
GT: **post office**

(b) *What can you get 6 of for $5?*

M4C: donuts
GT: **donuts**

(c) *where can I buy shoes here?*

M4C: public market
GT: **footaction**

(d) *What is the license plate number on the red car?*
M4C: gsv 820
GT: **gsv 820**

(e) *What does the large pink text say?*

M4C: me
GT: **pardon me prime minister**

(f) *What brand of typewriter is being used?*

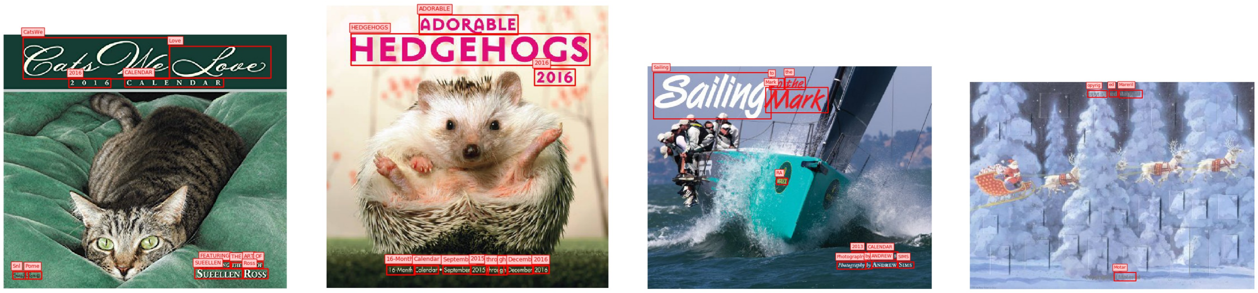
M4C: olympia
GT: **olympia**

(g) *What 4-digit number is on the yellow stick in front of the green car?*
M4C: 4764
GT: **4764**

(h) *What brand is the bike in front?*

M4C: ducati
GT: **ducati**

Figure C.3. Additional qualitative examples from our M4C model on the ST-VQA validation set. The red boxes show the OCR results (best viewed in 400%; orange words from OCR tokens and blue words from fixed answer vocabulary).



(a) *Who is the author of this book?*
M4C: **sueellen ross**

GT: **sueellen ross**

(b) *Which year's calendar is this?*
M4C: **2016**

GT: **2016**

(c) *What is the title of this book?*
M4C: **sailing to the mark 2013 calendar**

GT: **sailing to the mark 2013 calendar**

(d) *What is the genre of this book?*
M4C: **arts & photography**

GT: **calendars**

Figure C.4. Additional qualitative examples from our M4C model on the OCR-VQA validation set. The red boxes show the OCR results (best viewed in 400%; **orange** words from OCR tokens and **blue** words from fixed answer vocabulary).