Question-controlled Text-aware Image Captioning

Anwen Hu[1], Shizhe Chen[2], Qin Jin[1]

[1]AIM[3] Lab, School of Information, Renmin University of China

[2]INRIA

Outline

- Task Introduction
- Dataset Introduction
- Proposed Method
- Experiments
- Conclusion

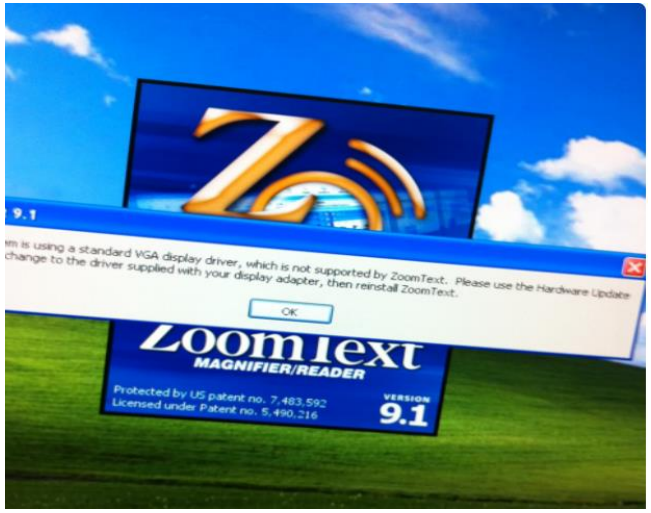
Task Introduction

Text-aware Image Captioning

TextCaps [1]

Ground Truth:
A sign in Spanish that says **Ruinas** and shows no pedestrians image.

VizWiz-Captions [2]



Ground Truth:
Computer screen displaying an error saying the display driver is not supported by **Zoom Text**.

scene text

[1] Gurari, Danna, et al. "Captioning Images Taken by People Who Are Blind." ECCV (17), 2020, pp. 417–434.
[2] Sidorov, Oleksii, et al. "TextCaps: A Dataset for Image Captioning with Reading Comprehension." ECCV (2), 2020, pp. 742–758.

Task Introduction

Text-aware Image Captioning (TextCap)

a book cover with the title uniunea europeana

author?

Task Introduction

Question-controlled Text-aware Image Captioning (Qc-TextCap)

a book

what is the title of the book?
Who wrote the book?

a book by gabriela carmen pascariu called uniunea europeana

Task Introduction

Question-controlled Text-aware Image Captioning (Qc-TextCap)

I

C^{ini}: a book

Q: what is the title of the book? Who wrote the book?

Y: a book by gabriela carmen pascariu called uniunea europeana

Input: <Image I, Initial Caption C^{ini}, Questions Q >
Output: Caption Y

Dataset Introduction

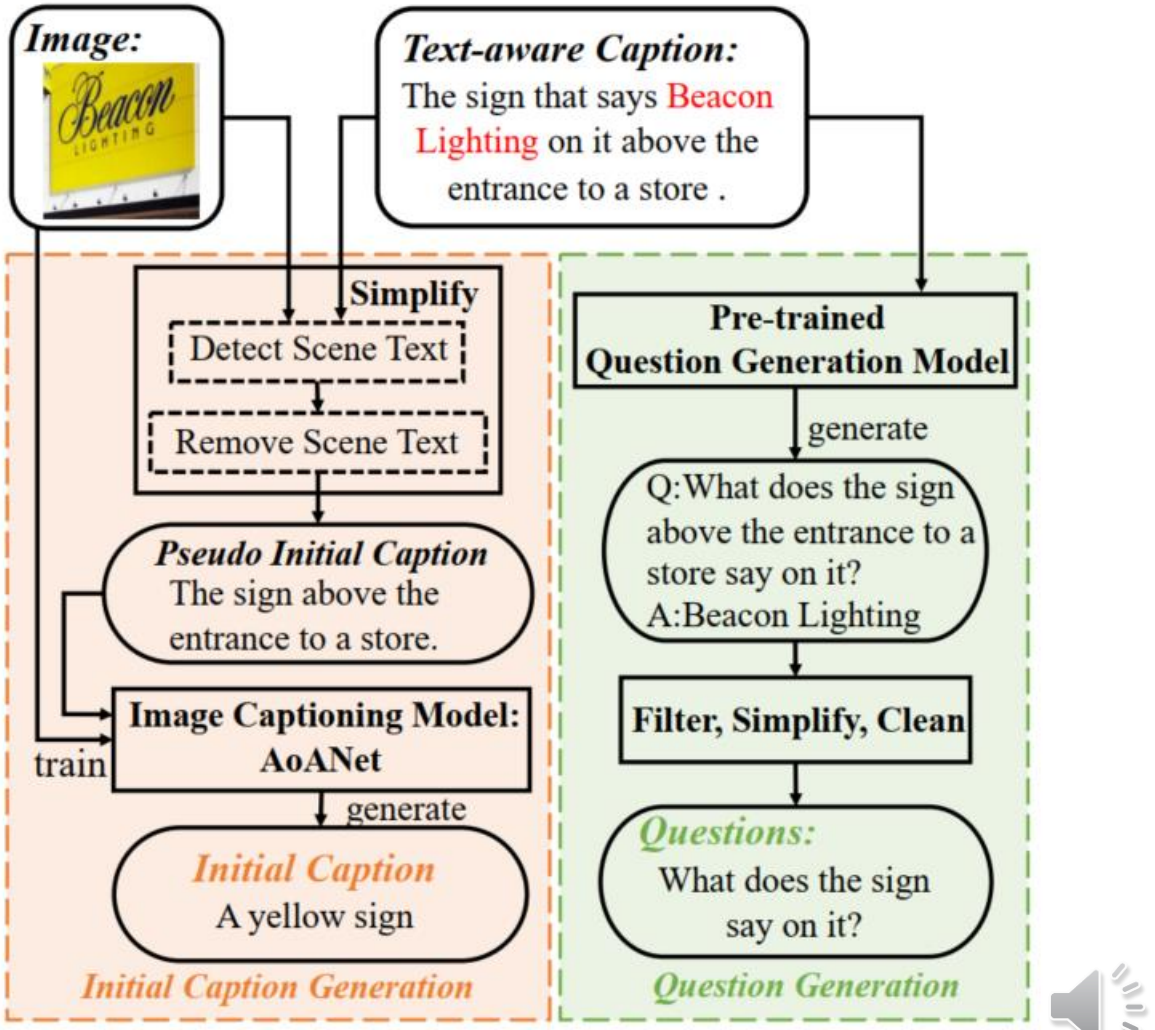
Automatic Dataset Construction

Initial Caption Generation:

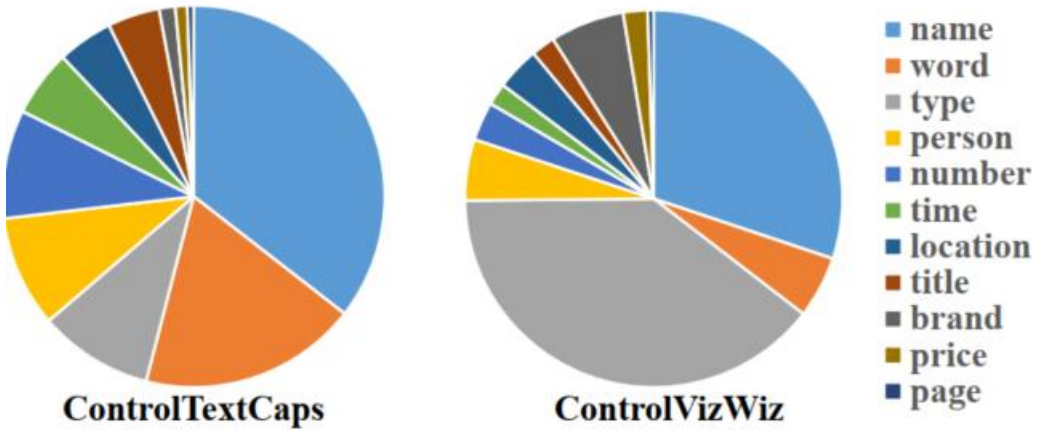
1. simplify text-aware captions, get pseudo initial captions.
2. train an in-domain general image captioning model
3. generate automatic initial captions.

Question Generation:

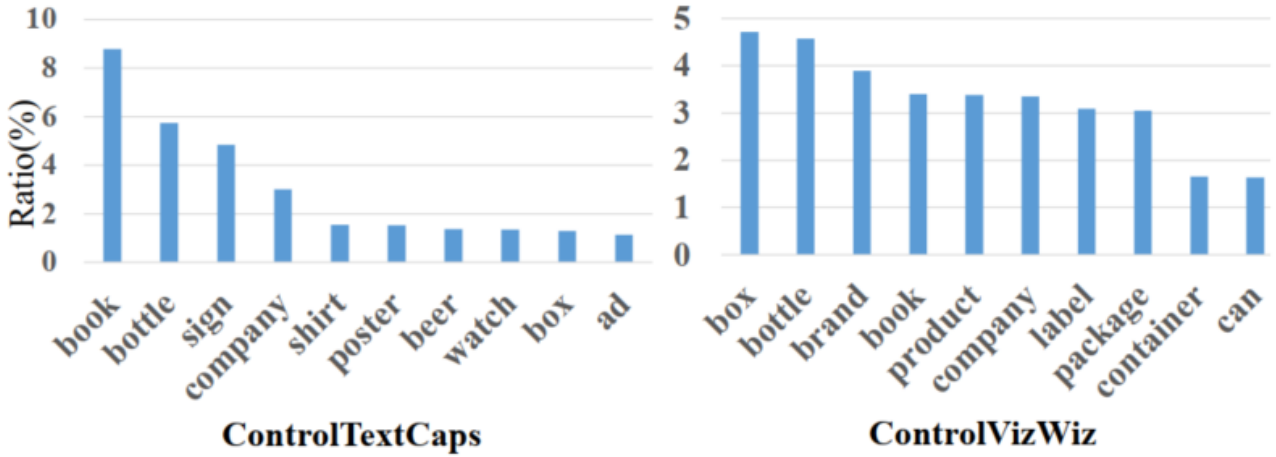
1. generate questions from text-aware captions
2. post process questions



Dataset Introduction

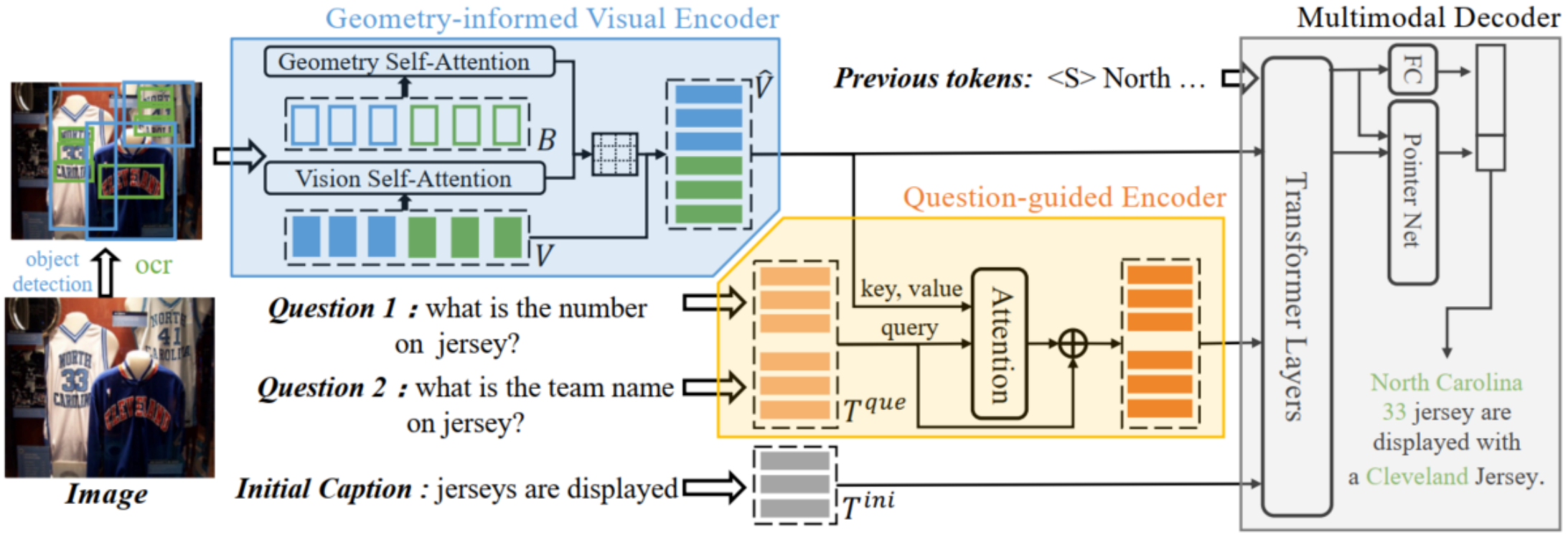


Question type distribution



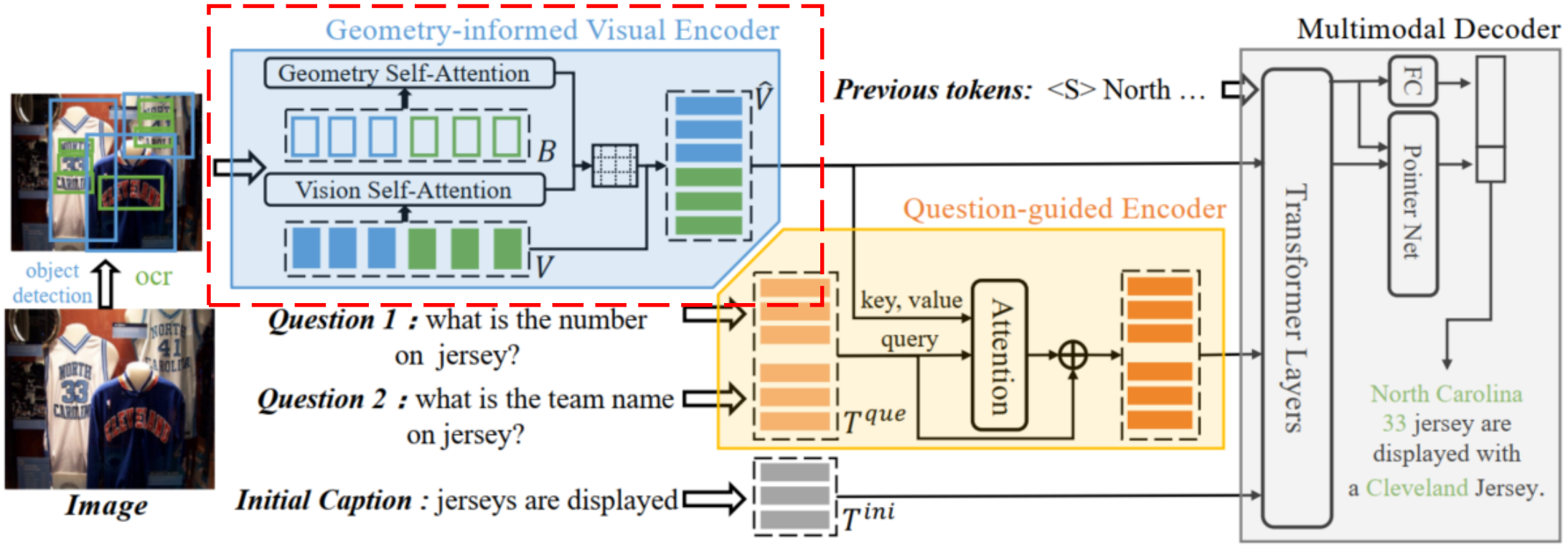
Top 10 objects in the questions of ‘name’ type

Proposed Method: GQAM



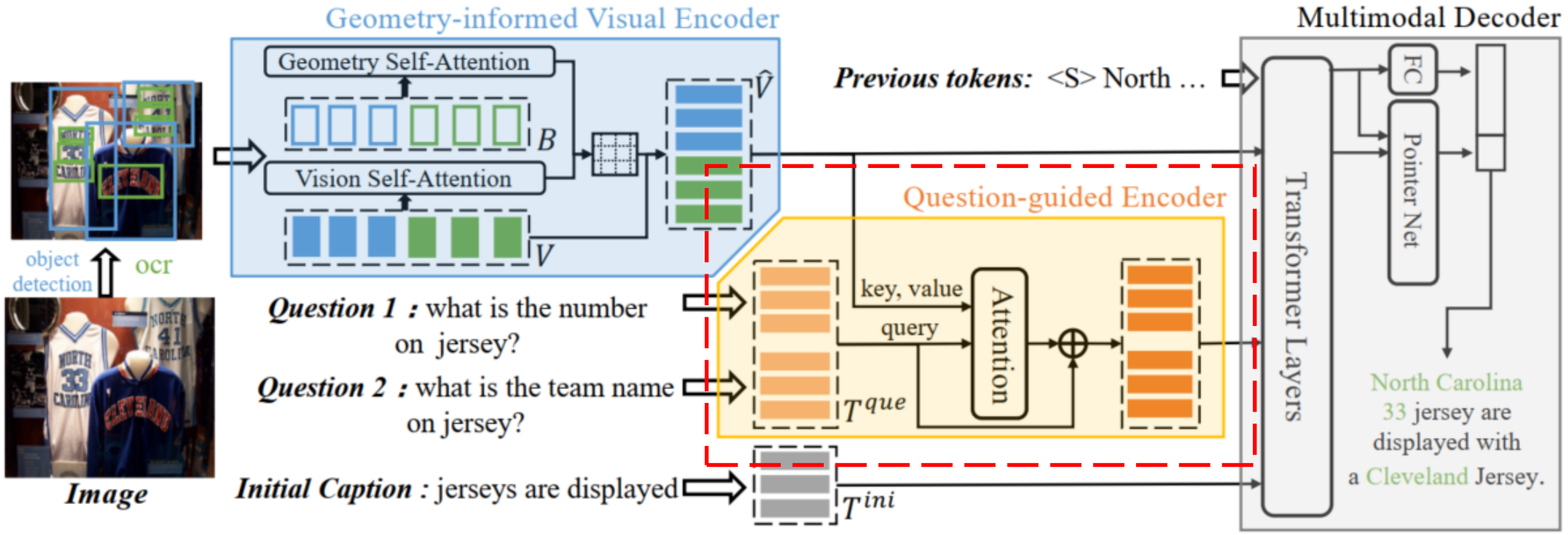
GQAM: Geometry and Question Aware Model

Proposed Method: GQAM



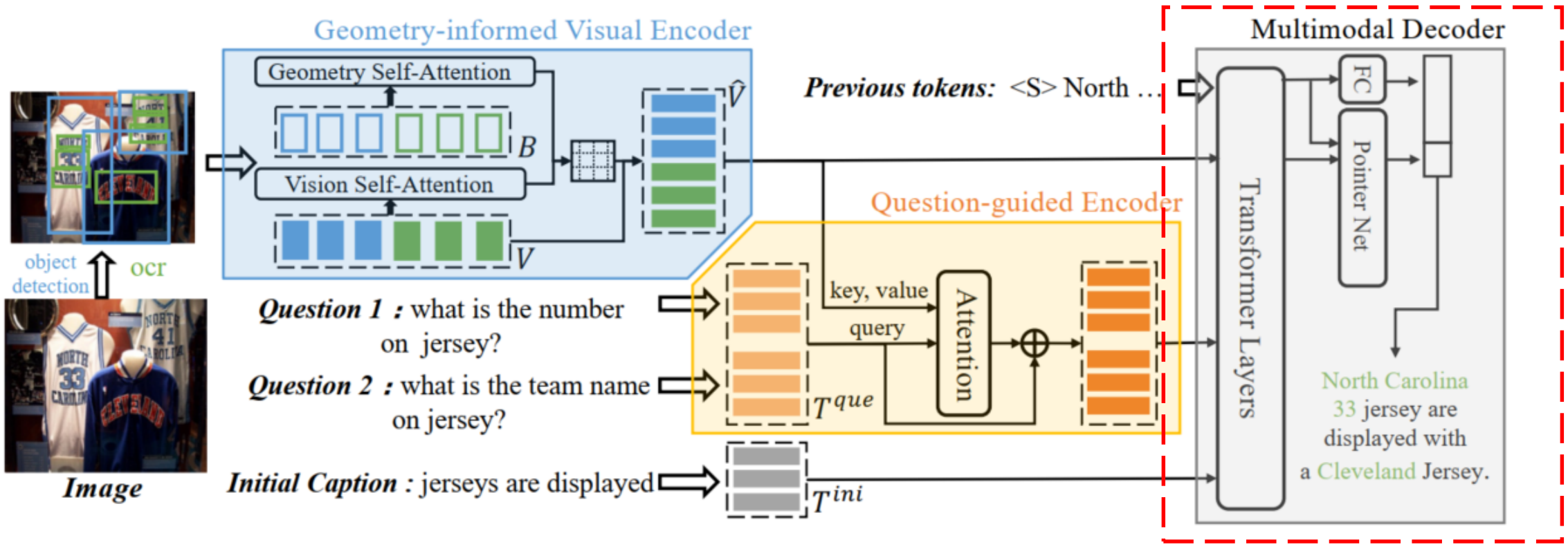
GQAM: Geometry and Question Aware Model

Proposed Method: GQAM



GQAM: Geometry and Question Aware Model

Proposed Method: GQAM



GQAM: Geometry and Question Aware Model

Experiments

Comparison of non-controllable and controllable models

Metrics:

***Overall Caption Quality*:** BLEU-n, METEOR, ROUGE-L, CIDEr, SPICE

Answering Ability: AnsRecall (the recall of answer tokens)

Table 2: Comparison of different models on the ControlTextCaps and ControlVizwiz datasets. 'Question' denotes whether the model takes questions as input.

Dataset	Model	Question	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr	SPICE	AnsRecall
ControlTextcaps	M4C-Captioner	✗	34.68	21.08	13.53	8.98	15.53	32.05	102.41	20.58	-
	ControlM4CC	✓	52.86	40.00	30.75	23.81	25.76	48.48	215.45	37.00	46.56
	GQAM w/o GE	✓	53.99	41.23	32.12	25.24	26.39	49.91	229.55	38.30	47.14
	GQAM	✓	**54.24**	**41.55**	**32.50**	**25.66**	**26.52**	**50.07**	**231.74**	**38.44**	**50.92**
ControlVizwiz	M4C-Captioner	✗	36.88	22.28	14.06	8.90	15.19	34.12	91.08	17.24	-
	ControlM4CC	✓	50.97	38.70	30.03	23.32	24.61	49.57	195.94	33.38	33.24
	GQAM w/o GE	✓	**53.00**	**40.67**	**31.90**	**25.03**	**25.25**	**50.55**	**210.60**	**34.58**	33.39
	GQAM	✓	51.61	39.62	31.06	24.33	24.82	49.73	201.35	33.81	**34.62**

Experiments

Comparison of different training strategies

I

pseudo initial caption :
The sign above the entrance to a store.

or

automatic initial caption:
A yellow sign

C^{ini}

What does the sign say on it?

Q

The sign that says Beacon Lighting on it above the entrance to a store .

Y

Table 3: Comparison of different training strategies. 'pseudo' or 'auto' means only using pesudo initial captions $\tilde{C}^{ini}$ or automatic initial captions C^{ini} as initial captions during training, respectively. 'rand(pesudo, auto)' means randomly choosing one of them for each training sample. During inference, only automatic initial captions are used as initial captions.

Dataset	Model	train strategy	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr	SPICE	AnsRecall
ControlTextcaps	GQAM	auto	54.24	41.55	32.50	25.66	26.52	50.07	231.74	38.44	50.92
		pesudo	43.26	29.39	20.74	14.72	19.89	38.97	143.36	25.46	49.47
		rand(auto, pseudo)	**54.63**	**42.01**	**32.96**	**26.13**	**26.83**	**50.50**	**238.20**	**38.69**	**51.27**
ControlVizwiz	GQAM	auto	53.00	40.67	31.90	25.03	25.25	50.55	210.60	34.58	33.39
		pseudo	44.85	30.56	21.70	15.67	20.01	41.60	140.08	23.77	**34.70**
		rand(auto, pseudo)	**54.41**	**42.43**	**33.64**	**26.79**	**25.98**	**51.65**	**223.23**	**35.85**	33.72

Experiments

Table 4: Diversity evaluation of our GQAM and the text-aware captioning model M4C-Captioner.

Dataset	Model	Div-1	Div-2	SelfCIDEr
ControlTextCaps	M4C-Captioner	7.44	21.11	62.58
	GQAM	**14.72**	**38.00**	**78.32**
ControlVizWiz	M4C-Captioner	6.41	19.97	56.36
	GQAM	**10.88**	**28.71**	**63.06**

Table 5: Human evaluation of accurate scene text information (ST Info) and overall caption quality. For simplicity, we use M4CC to refer to M4C-Captioner

Dataset		ST Info	Overall Quality
ControlTextcaps	GQAM>M4CC	43.48%	51.38%
	GQAM≈M4CC	42.29%	27.67%
	GQAM<M4CC	14.23%	20.95%
ControlVizWiz	GQAM>M4CC	44.30%	41.77%
	GQAM≈M4CC	39.24%	24.05%
	GQAM<M4CC	16.46%	34.18%

Experiments

M4C-Captioner:
a book cover with the title uniunea europeana

Ground A: a book by gabriela carmen pascariu about uniunea europeana

(a)

Initial caption: a book
----------***question-controlled text-aware captions***----------
Questions A: what is the title of the book? who wrote the book?
ControlM4CC: a book by gabriela carmen pascariu called uniunea europeana
GQAM: a book by gabriela carmen pascariu called uniunea politici si piete europeana

M4C-Captioner:
a bottle of the royal legac sits on a table

Ground A: a bottle of Royal Legacy malt whiskey and the box it came in.
Ground B: bottle of alcohol that says The Royal Legacy by a green box.

(b)

Initial Caption: a bottle next to a box
----------***question-controlled text-aware captions***----------
Questions A: what is the brand on the bottle? what is in the bottle?
ControlM4CC: a bottle of royal royal legac next to a box of it
GQAM: a bottle of royal legac malt whisky liqueur next to a box of the new Orleans

Questions B: what does the label on the bottle say? what is in the bottle?
ControlM4CC: a bottle of alcohol that says the royal legac on it
GQAM: a bottle of beer that says " royal legacy " is on the table

M4C-Captioner:
a phone that has the word mil. at&t on it

Ground A: white phone with a screen that says August 12th on it.
Ground B: a mobile phone using AT&T's network shows an app on its screen that is used to monitor baby feeding times and amounts.

(c)

Initial Caption: a white phone on a wooden table
----------***question-controlled text-aware captions***----------
Questions A: what is the date shown on the phone?
ControlM4CC: a white phone on a table that says ' sunday august 12 2012 ' on it
GQAM: a phone on a wooden table with the date of sunday august 12 2012

Questions B: what app is installed on the phone?
ControlM4CC: a phone with the app feeding on the screen
GQAM: a phone with the app feeding and sleeping on the screen

Qualitative results

Conclusions

- To generate personalized text-aware captions, we define a challenging task, namely Question-controlled Text-aware Image Captioning (Qc-TextCap).
- We develop an automatic system to construct two appropriate datasets based on existing TextCaps and VizWiz-Captions datasets.
- We propose a Geometry and Question Aware Model (GQAM) for Qc-TextCap.
- Codes and datasets are available at https://github.com/HAWLYQ/Qc-TextCap

THANK YOU

If any questions, feel free to contact Anwen Hu
Email: anwenhu@ruc.edu.cn

中国人民大學
RENMIN UNIVERSITY OF CHINA