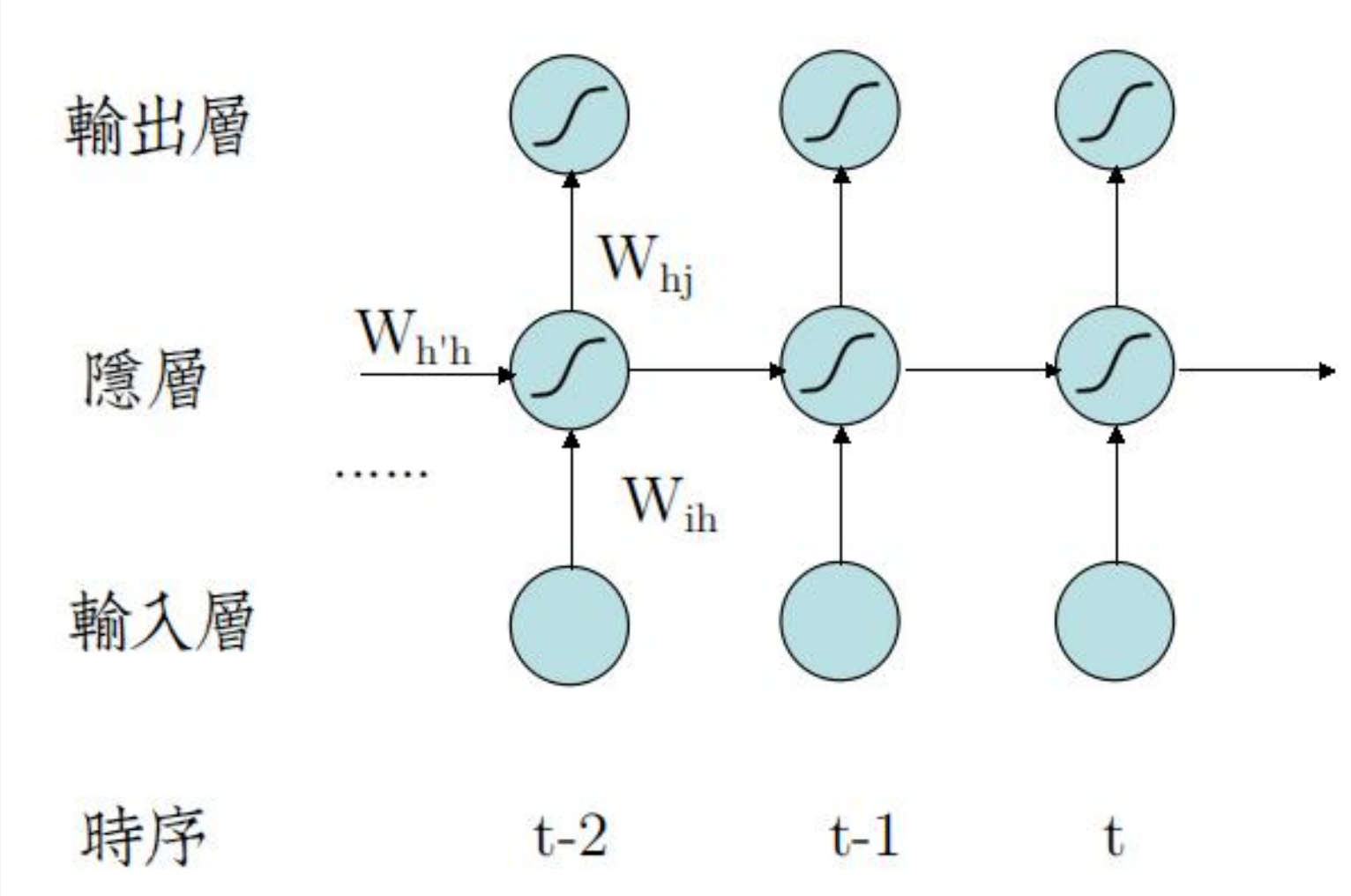
IMAGE & SENTENCE

PAPER SHARE BY WQX

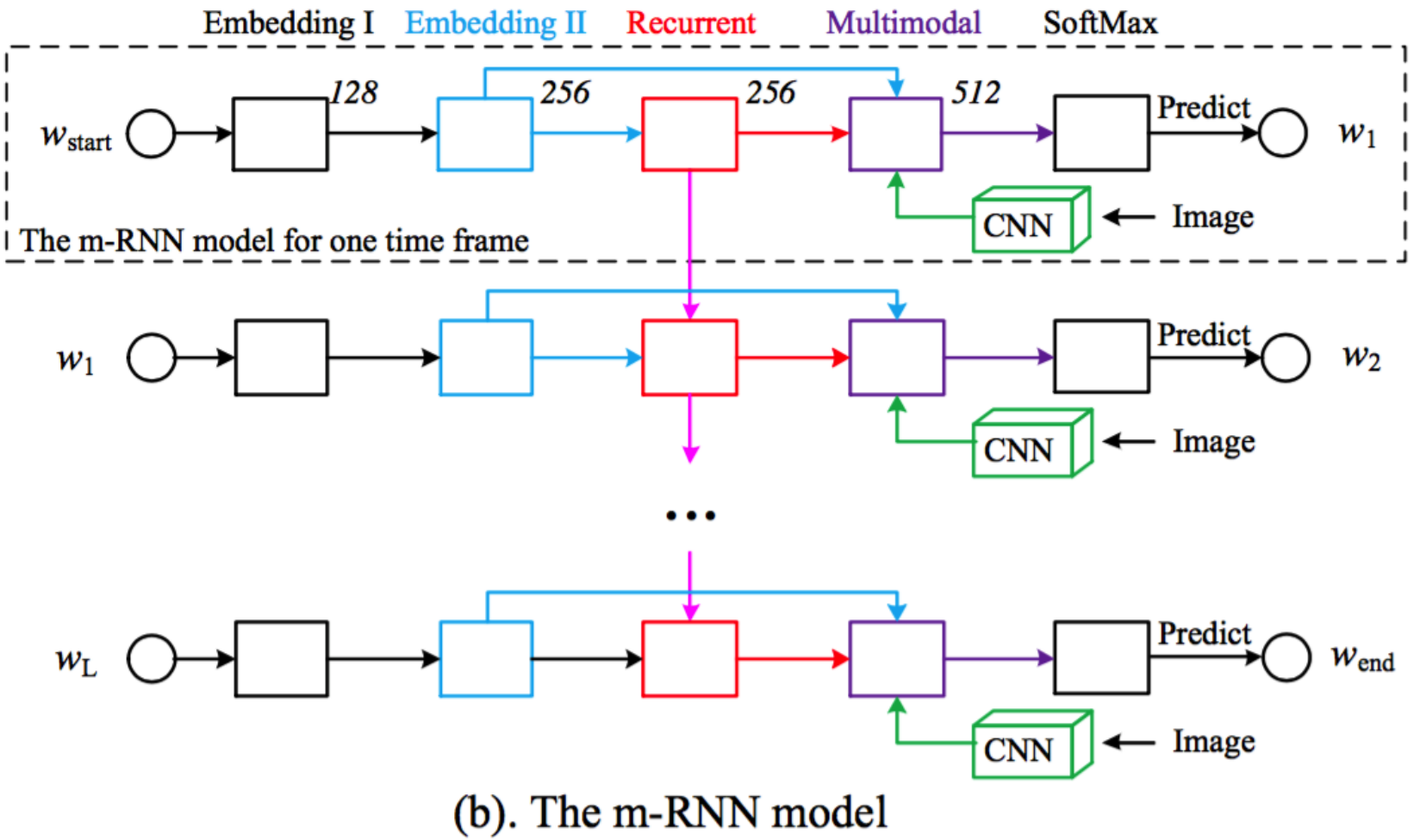
DEEP CAPTIONING WITH MULTIMODAL RECURRENT NEURAL NETWORKS (M-RNN)

ARE YOU TALKING TO A MACHINE?
DATASET AND METHODS FOR MULTILINGUAL IMAGE QUESTION ANSWERING

RNN (LSTM) IMAGE CAPTION



BAIDU'S NEW MODEL



(b). The m-RNN model

ALGORITHMS

model part and the vision part of the m-RNN model (see Figure 2(b)). This layer has three inputs: the word-embedding layer II, the recurrent layer and the image representation. For the image representation, here we use the activation of the 7th layer of AlexNet (Krizhevsky et al. (2012)) or 15th layer of VggNet (Simonyan & Zisserman (2014)), though our framework can use any image features. We map the activation of the three layers to the same multimodal feature space and add them together to obtain the activation of the multimodal layer:

$$\mathbf{m}(t) = g_2(\mathbf{V}_w \cdot \mathbf{w}(t) + \mathbf{V}_r \cdot \mathbf{r}(t) + \mathbf{V}_I \cdot \mathbf{I}); \quad (3)$$

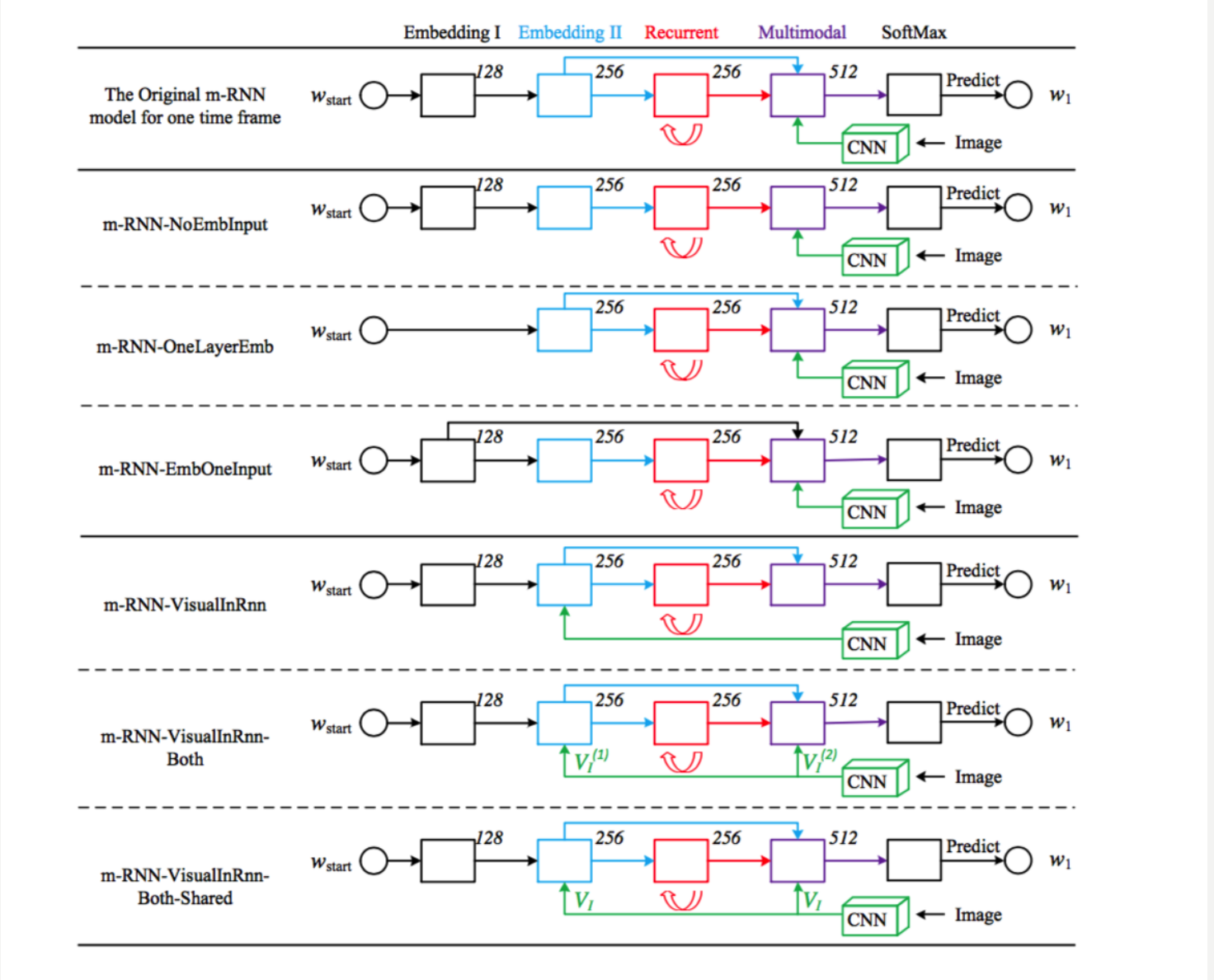
where "+" denotes element-wise addition, $\mathbf{m}$ denotes the multimodal layer feature vector, $\mathbf{I}$ denotes the image feature. $g_2(.)$ is the element-wise scaled hyperbolic tangent function (LeCun et al. (2012)):

$$g_2(x) = 1.7159 \cdot \tanh(\frac{2}{3}x) \quad (4)$$

This function forces the gradients into the most non-linear value range and leads to a faster the training process than the basic hyperbolic tangent function.

$$\mathcal{C} = \frac{1}{N}\sum_{i=1}^{N_s} L_i \cdot \log_2 \mathcal{PPL}(w_{1:L}^{(i)} | \mathbf{I}^{(i)}) + \lambda_\theta \cdot \|\theta\|_2^2$$

MODEL COMPARE



RESULTS

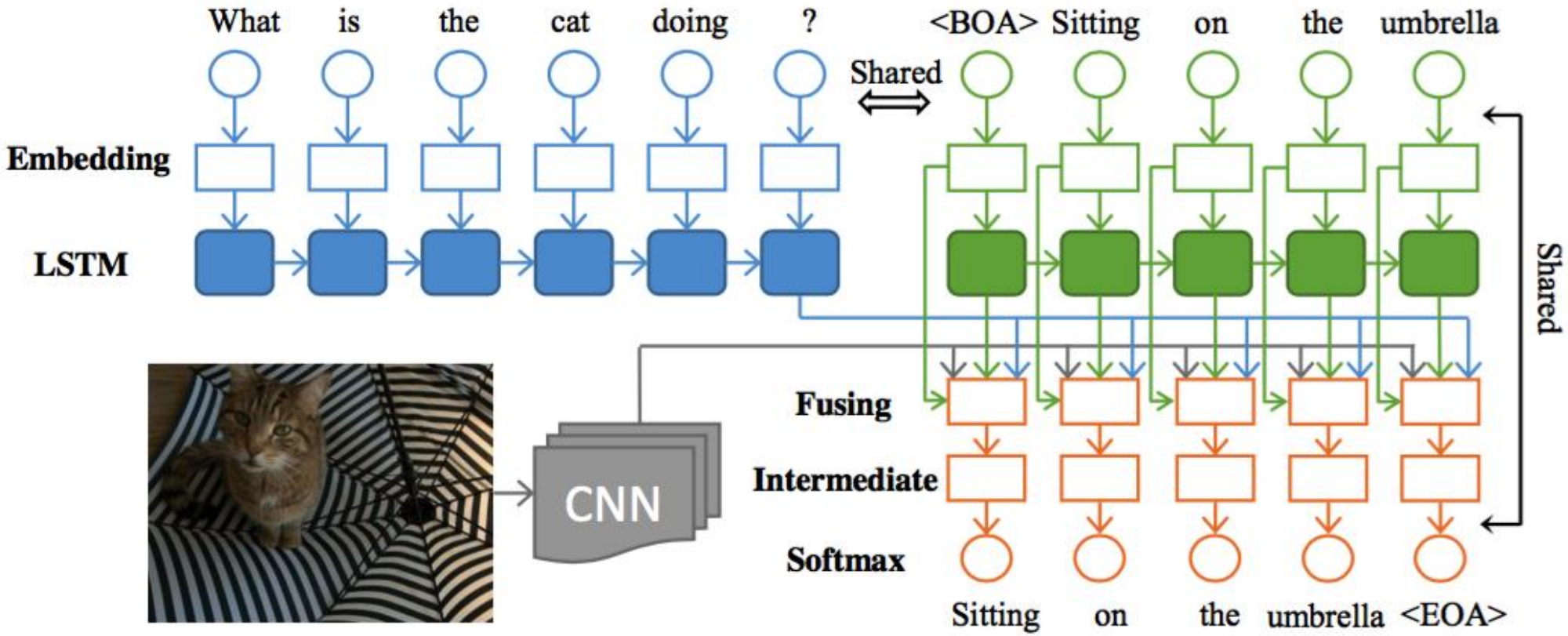
1. Tourists are sitting at a long table with beer bottles on it in a rather dark restaurant and are raising their bierglaeser;
2. Tourists are sitting at a long table with a white table-cloth in a somewhat dark restaurant;

Tourists are sitting at a long table with a white table cloth and are eating;

1. A dry landscape with light brown grass and green shrubs and trees in the foreground and large reddish-brown rocks and a blue sky in the background;
2. A few bushes at the bottom and a clear sky in the background;

A dry landscape with green trees and bushes and light brown grass in the foreground and reddish-brown round rock domes and a blue sky in the background;

M-QA


What
is
the
cat
doing
?
<BOA>
Sitting
on
the
umbrella
Shared
Embedding
LSTM
Shared
CNN
Fusing
Intermediate
Softmax
Sitting
on
the
umbrella
<EOA>

DATASET

Image			
Question	公共汽车是什么颜色的？ What is the color of the bus?	黄色的是什么？ What is there in yellow?	草地上除了人以外还有什么动物 What is there on the grass, except the person?
Answer	公共汽车是红色的。 The bus is red.	香蕉。 Bananas.	羊。 Sheep.

MORE

戴帽子的男孩在干什么？
What is the boy in green cap doing?

他在玩滑板。
He is playing skateboard.

图片中有人么？
Is there any person in the image?

有。
Yes.

电脑在老人的左面还是右面？
Is the computer on the right hand or left hand side of the gentleman?

右手侧。
On the right hand side.

飞盘是什么颜色？
What is the color of the frisbee?

黄色。
Yellow.

公交车停在那干吗？
Why does the bus park there?

准备维修。
Preparing for repair.

房间里的沙发是什么质地的？
What is the texture of the sofa in the room?

布艺。
Cloth.

这个人在挑菜么？
Is the man trying to buy vegetables?

是的。
Yes.

这个蛋糕是几层的？
How many layers are there for the cake?

六层。
Six.

这些人在做什么？
What are the people doing?

打雨伞步行。
Walking with umbrellas.

手机，鼠标，电脑混放表示什么？
What does it indicate when the phone, mouse and laptop are placed together?

主人困了，睡着了
Their owner is tired and sleeping.

SUCCESSFUL CASE

这是在什么地方？
Where is this?

这是在厨房。
This is the kitchen room.

这个人在打网球么？
Is this guy playing tennis?

是的。
Yes.

这是什么食物？
What kind of food is this?

披萨。
Pizza.

电脑在哪里？
Where is the computer?

在桌子上。
On the desk.

Figure 5: The sample generated questions by our model and their answers.

FAILED CASE

帅哥在干什么？
What is the handsome boy doing?

在抓飞盘。
Trying to catch the frisbee.

冲浪 。
Surfing.

这是什么？
What is there in the image?

草原上的马群 。
Horses on the grassland.

这是牛。
They are buffalos.

盘子里有什么水果？
Which fruit is there in the plate?

苹果和橙子 。
Apples and oranges.

香蕉和橙子 。
Bananas and oranges.

这是什么车？
What is the type of the vehicle?

公交汽车。
Bus.

火车。
Train.

公交车停在那干吗？
Why does the bus park there?

准备维修。
Preparing for repair.

<OOV>。
<OOV> (I do not know.)

Figure 6: Failure cases of our mQA model on the FM-IQA dataset.

PART RIGHT CASE

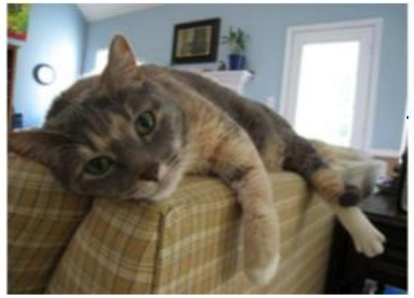
Image					
Question	盘子里有什么？ What is in the plate?	狗在干嘛？ What is the dog doing?	小猫在哪里？ Where is the cat?	这是什么？ What is there in the image?	这是什么车？ What is the type of the vehicle?
Answer	食物。 food.	在冲浪。 Surfing in the sea.	床上。 On the bed.	这是钟表。 There is a clock.	火车 Train.

Figure 4: Random examples of the answers generated by the mQA model with score “1” given by the human judges.

RESULTS

	Visual Turing Test			Human Rated Scores			
	Pass	Fail	Pass Rate (%)	2	1	0	Avg. Score
Human	948	52	94.8	927	64	9	1.918
blind-QA	340	660	34.0	-	-	-	-
mQA	647	353	64.7	628	198	174	1.454

Table 1: The results of our mQA model for our FM-IQA dataset.

THANK YOU!