

# Diversity and Consistency: Exploring Visual Question-Answer Pair Generation

Sen Yang<sup>1</sup>, Qingyu Zhou<sup>2</sup>, Dawei Feng<sup>1</sup>, Yang Liu<sup>1</sup>, Chao Li<sup>2</sup>,  
Yunbo Cao<sup>2</sup>, Dongsheng Li<sup>1\*</sup>

<sup>1</sup>Computer College, National University of Defense Technology

<sup>2</sup>Tencent Cloud Xiaowei

yangsen@nudt.edu.cn, qingyuzhou@tencent.com

davyfeng.c@gmail.com, lds1201@163.com

## Abstract

Although showing promising values to downstream applications, generating question and answer together is under-explored. In this paper, we introduce a novel task that targets question-answer pair generation from visual images. It requires not only generating diverse question-answer pairs but also keeping the consistency of them. We study different generation paradigms for this task and propose three models: the pipeline model, the joint model, and the sequential model. We integrate variational inference into these models to achieve diversity and consistency. We also propose region representation scaling and attention alignment to improve the consistency further. We finally devise an evaluator as a quantitative metric for consistency. We validate our approach on two benchmarks, VQA2.0 and Visual-7w, by automatically and manually evaluating diversity and consistency. Experimental results show the effectiveness of our models: they can generate diverse or consistent pairs. Moreover, this task can be used to improve visual question generation and visual question answering.

## 1 Introduction

Teaching a machine to generate question-answer pairs (QAPs) from images can benefit a lot of downstream applications such as child education (Wang et al., 2018), visual dialog (Das et al., 2017), generating verification code for websites, visual question generation (VQG) (Krishna et al., 2019) and visual question answering (VQA) (Wu et al., 2016). For example, training VQA models requires large scale labelled data, which is usually labour intensive and expensive to construct. Meanwhile, bias still exists in large QA datasets (Goyal et al., 2017), including domain coverage, question and answer types, and

\*Corresponding author.



Images	Questions	Answers
	What is this? Where is cake? Why is the cake there? What color is on bottom of cake?	Birth day cake. In box. Birth day party. Pink.
Images	Questions	Answers
	What is on the computer? What color is the keyboard? Is the operator left or right handed? What color is the computer on the left?	Website White Right Black

Figure 1: Two instances of VQAPG. The input contains only an image, while the target contains both question and answer.

linguistic style. Therefore, as an alternative to constructing datasets manually, QAP generation can promote VQA further.

In this paper, we first study a novel task that aims to generate QAPs from visual images, namely VQAPG (Visual Question Answer Pair Generation). As shown in Figure 1, it has two challenges: diversity and consistency. On the one hand, even a simple image will contain various content that could be asked. The focus of questions could range from appeared objects and their features to the global attributes, such as the image background and photo time. On the other hand, generating consistent QAPs is also critical. When we say a QAP is consistent with the image, we mean two points: the question is answerable with the image, and the answer is correct for the question. Therefore, keeping the consistency is difficult because it requires the generation model to simultaneously guarantee the correctness of both questions and answers.

There are two related tasks with VQAPG, but neither serves as suitable prior art. The first is VQG (Zhang et al., 2017; Patro et al., 2018; Krishna et al., 2019), which produces questions given answers or other knowledge. Another similar task is VQA that aims to answer given questions (Goyal et al., 2017; Zhou et al., 2020; Su et al., 2020; Lu et al., 2019). They can be viewed as two subtasks

of VQAPG. However, simply combining the two is not an ideal substitute for VQAPG since they condition another’s output for learning.

We study several generation paradigms to perform VQAPG. The first is a pipeline model that generates questions and answers one after another. Next, inspired by the non-autoregressive text generation (Ren et al., 2020), we propose a joint model that generates questions and answers in parallel. To reduce the model size, we also propose a sequential model that concatenates the two targets into one. We integrate latent variable(s) into these models through variational inference (Kingma and Welling, 2014) to improve diversity. If fed with different latent variables sampled from the prior distribution, the model can generate various QAPs. Besides, we observe that if we grid an image into multiple regions, the target question and answer are only related tightly with only part of them. As the latent variable contains information of the target QAP, we use it to make the model concentrate on those related regions by scaling their representations. To improve the joint model’s consistency, we align the attentions of the question decoder and the answer decoder to make them focus on similar regions.

A remaining issue of VQAPG is how to measure the consistency automatically in addition to manual inspection. Traditional popular metrics, such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005), only tell the overlapping degree between generated QAPs and the target ones, being insufficient to indicate the consistency. Targeting this issue, we devise a consistency evaluator trained with an adversarial strategy. We assume the samples in the dataset as consistent and construct inconsistent samples through shuffling images, answers, questions, or all of them. This evaluator will give a high score if the generated QAP is consistent with the image and a low score if not.

We conduct experiments on two datasets, VQA2.0 (Goyal et al., 2017) and Visual-7w (Zhu et al., 2016), in both of which each image has multiple human-created QAPs. The quantitative results indicate that each model has its strengths. The pipeline model achieves the best diversity, while the sequential model achieves the best consistency. However, they can not boost both together. In contrast, the joint model improves both diversity and consistency impressively through variation. More ablation studies illustrate the effectiveness of our

proposed methods: both the region representation scaling mechanism and the attention alignment mechanism improve consistency significantly. We also evaluate generated QAPs manually and give case studies. Moreover, to prove the effectiveness of VQAPG on downstream applications, we use VQAPG to generate pre-training samples for VQG and VQA. Results indicate VQAPG can enhance the performance of both VQG and VQA models.

In short, our contribution mainly includes four parts: **i)** We propose a novel task, VQAPG, that targets to generate diverse and consistent question-answer pairs from images. **ii)** To perform VQAPG, we study multiple generation paradigms and propose three models. We also design a consistency evaluator. **iii)** We incorporate variational inference into models and propose a series of techniques, including region representation scale and attention concentration to improve diversity and consistency. **iv)** We conduct comprehensive experiments on two large scale datasets. The results show the effectiveness of our approach to generate diverse and consistency QAPs and benefit other applications.

## 2 Related Works

**Visual Question Generation** is an interesting task emerged in recent years. Question generation is firstly studied on text (Heilman and Smith, 2010; Labutov et al., 2015; Du et al., 2017; Zhou et al., 2018; Sun et al., 2018; Ma et al., 2020; Kim et al., 2019). While related studies on images has received little attention. Existing methods in this field are typically based on learning algorithms (Mostafazadeh et al., 2016; Zhang et al., 2017). Such methods are often incorporated with the variational process (Jain et al., 2017; Krishna et al., 2019). Visual question generation is also conducted together with visual question answering (Li et al., 2018; Sun et al., 2020).

**Visual Question Answering** has received more interest thanks to available public datasets such as VQA2.0 (Goyal et al., 2017) and VisualGenome (Krishna et al., 2017). Through fine-tuning on large pre-trained models (Zhou et al., 2020; Su et al., 2020; Lu et al., 2019), the performance has been improved considerably. However, it still requires large scale labeled datasets, which is too consuming to annotate manually. Therefore, a successful VQAPG system would be beneficial to reduce such costs.

**Question Answer Pair Generation** on images

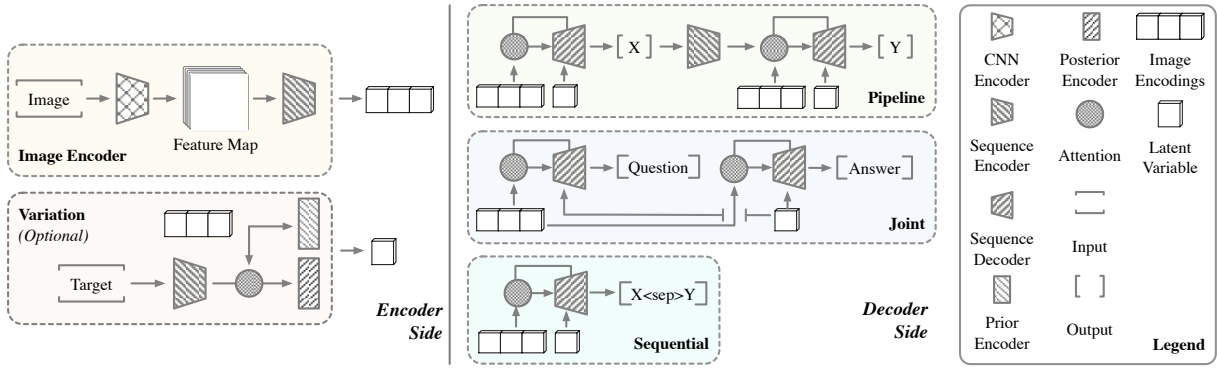


Figure 2: Overall architecture of our three models for VQAPG. They share the same image encoder architecture. The variation module is optional to produce latent variables. Here, “X” and “Y” represent question and answer, or vice visa. “Target” represents the sequence for estimating posterior distribution, such as question, answer, or both. “<sep>” is a reserved token to separate question and answer.

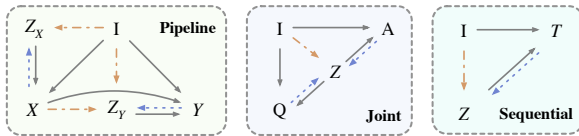


Figure 3: Three graphical models. The orange dashed line indicates the influence works for estimating both prior and posterior. The blue dashed line indicates the influence works only for estimating posterior. The solid gray line indicates the influence works for estimating likelihood.  $T$  represents “ $X <sep> Y$ .”

is unexplored so far. Some works explore this task in text using techniques such as pipeline (Subramanian et al., 2018), multi-agent system (Wang et al., 2019), hierarchical variational model (Subramanian et al., 2018) or coreference knowledge (Lee et al., 2020). Su et al. (2021) also proposes a model for QAP generation from video. However, such QAP generation works assume answers are selected from the spans of input context (Subramanian et al., 2018; Lee et al., 2020; Wang et al., 2019) or the given candidates (Su et al., 2021). As answers could not be extracted directly from images and there are no candidate ones, the above methods can not be simply applied to the image.

### 3 VQAPG Task

Given an image denoted as  $I$ , VQAPG aims to produce diverse QAPs, each of which contains an answerable question  $Q$  and its correct answer  $A$  under  $I$ . Both the question  $Q = q_1, q_2, \dots, q_m$  and the answer  $A = a_1, a_2, \dots, a_n$  are sequences, in which  $q_i$  and  $a_i$  represent tokens. Figure 4 compares VQAPG with typical VQG and VQA task. Our final goal is to obtain a model to approximate

the true data distribution  $P(Q, A|I)$  so that we can sample questions and answers from it. Because VQAPG is a one-to-many task, we also expect this model to support sampling of diverse and consistent QAPs.

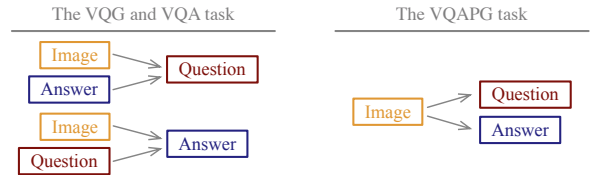


Figure 4: Comparison between VQG&VQA and VQAPG. Unlike VQG or VQA, which rely on answer or question additionally for a generation, VQAPG generates question and answer both only in the presence of image.

## 4 Approach

We propose three models to perform VQAPG, including the pipeline model, the joint model, and the sequential model. As shown in Figure 2, all of them adopt the encoder-decoder architecture.

### 4.1 Preliminary

All three models rely on the same image encoder to embed images into vector space. It contains a convolutional neural network (CNN) and a context-aware sequence model. The former transforms the image into a feature map of size  $C \times H \times W$ , which can be viewed as embeddings of  $R^1$  regions with dimension  $C$ . The latter encodes further the feature map. We use  $\mathbf{h}_I = (h_1, \dots, h_R)$  to represent the final image representation, in which  $h_i$  is the  $i$ -th region representation.

<sup>1</sup>For brevity, we represent  $H \times W$  as  $R$ .

Model	Log Likelihood	ELBO
Pipeline	$\log P_{\Omega_X}(X I)$	$\mathbb{E}_{Z_X} \log P_{\Omega_X}(X I, Z_X) - \text{KL}(P_{\Psi_X}(Z_X I, X)  P_{\Phi_X}(Z_X I))$
	$\log P_{\Omega_Y}(Y I, X)$	$\mathbb{E}_{Z_Y} \log P_{\Omega_Y}(Y I, X, Z_Y) - \text{KL}(P_{\Psi_Y}(Z_Y I, X, Y)  P_{\Phi_Y}(Z_Y I, X))$
Joint	$\log [P_{\Omega_Q}(Q I)P_{\Omega_A}(A I)]$	$\mathbb{E}_Z \log [P_{\Omega_Q}(Q I, Z)P_{\Omega_A}(A I, Z)] - \text{KL}(P_{\Psi}(Z I, Q, A)  P_{\Phi}(Z I))$
Sequential	$\log P_{\Omega}(T I)$	$\mathbb{E}_Z \log P_{\Omega}(T I) - \text{KL}(P_{\Psi}(Z I, T)  P_{\Phi}(Z I))$

Table 1: The objectives of the three models. "KL" means the Kullback-Leibler distance. The pipeline contains two sub-models, and they are trained with different objectives.

Each model has two versions, the baseline and the variational. The goal of the baseline is to maximize the conditioned likelihood, which is estimated by  $P_{\Omega}$ , where  $\Omega$  denotes parameters of the model. Table 1 summarizes objectives of all models.

However, only maximizing the likelihood is insufficient to generate diverse QAPs, so we incorporate variation inference (Kingma and Welling, 2014) (as shown in Figure 3). The key is to estimate the prior and posterior distributions of the latent variable represented as  $Z$ . We define both distributions as isotropic Gaussian and design a prior encoder and a posterior encoder for estimation. We use the symbol  $\Phi$  to represent parameters of the prior encoder and  $\Psi$  for the posterior encoder. Both of them use a multi-layer perceptron to produce the mean and variance of distributions. We exploit reparametrization trick (Kingma and Welling, 2014) to allow backpropagation of gradients. With latent variables, the objective of variational models is to maximize the evidence lower bound (ELBO) as shown in Table 1 (The derivation details of ELBO are shown in Appendix.).

Since one image contains multiple regions and the target QAP will focus only a few of them, we argue that we should allocate different weights to these regions and encode them into the latent variable. So we exploit a text-image attentional module for both prior and posterior encoder. Given the image encoding  $\mathbf{h}_I$  and a sequence encoding  $\hat{h}$ , the attention result is:

$$\begin{aligned}
 f(\hat{h}, \mathbf{h}_I) &= \mathbf{M}(\hat{h}||h') \\
 h' &= \alpha^\top h_i \\
 \alpha &= \text{Softmax}(\hat{h}^\top \mathbf{h}),
 \end{aligned} \tag{1}$$

where  $\mathbf{M}$  represents a multi-layer perceptron.

## 4.2 Pipeline Model

The pipeline model contains two sub-models that conduct question generation and answer generation separately. The generation order of question and

answer is changeable. Therefore, we use  $X$  to represent the first generated and  $Y$  for the second and use  $X$ -model and  $Y$ -model to represent two corresponding sub-models, respectively. The  $Y$ -model takes gold  $X$  as input during training, and predicted  $X$  produced by  $X$ -model during inference.

In the baseline version, the two sub-models just aim to maximize the likelihood, as shown in Table 1. In the variational version, we add two latent variables,  $Z_X$  and  $Z_Y$ . The former is used to control the diversity of  $X$  and the latter is for  $Y$ . As shown in Figure 3, the prior distribution and posterior of  $Z_X$  are estimated as:

$$P_{\Phi_X}(Z_X|I) = \mathbf{M}\left(\frac{1}{R} \sum_{i \leq R} h_i\right) \tag{2}$$

$$P_{\Psi_X}(Z_X|I, X) = \mathbf{M} \circ f(h_X, \mathbf{h}_I),$$

where  $h_X$  is the representation of  $X$  obtained from a sequence encoder, " $\circ$ " represents composition. For  $Z_Y$ , the prior and posterior are:

$$\begin{aligned}
 P_{\Phi_Y}(Z_Y|I, X) &= \mathbf{M} \circ f(h_X, \mathbf{h}_I) \\
 P_{\Psi_Y}(Z_Y|I, X, Y) &= \mathbf{M} \circ f(h_X \oplus h_Y, \mathbf{h}_I),
 \end{aligned} \tag{3}$$

where  $h_Y$  is the representation of  $Y$ , " $\oplus$ " indicates concatenation.

## 4.3 Joint Model

Inspired by the non-autoregressive text generation (Ren et al., 2020), we propose a joint model that generates questions and answers in parallel with two decoders as shown in Figure 2.

The joint model assumes  $Q$  and  $A$  are independent conditioned on  $I$  and optimizes two decoders separately without considering each other. As an image could map to multiple QAPs, the consistency of generated QAP can not be guaranteed in the baseline joint model. On the other hand, the latent variable  $Z$  contains the information of the target QAP. Therefore, the consistency can be improved by introducing  $Z$ . As shown in Figure 3,

the prior of  $Z$  is estimated as  $P_\Phi(Z|I)$ , alike with  $P_{\Phi_X}(Z_X|I)$  in Equation 2, and the estimated posterior  $P_\Psi(Z|I, Q, A)$  is alike with  $P_{\Psi_Y}(Z_Y|I, X, Y)$  in Equation 3.

To improve the consistency further, we argue that the two decoders in the joint model should focus on similar regions for better consistency. Therefore, we align the attention to regions used in the two decoders. Specifically, we add an attention alignment term  $\mathcal{L}_{attn}$  to the objective of the joint model:

$$\mathcal{L}_{attn} = -\lambda \text{KL}\left(\frac{1}{m} \sum_{i \leq m} \beta_i \parallel \frac{1}{n} \sum_{i \leq n} \gamma_i\right). \quad (4)$$

$\beta_i$  is the attention of the question decoder at time step  $i$  and  $\gamma_i$  is the attention of the answer decoder. They are computed alike with  $\alpha$  in Equation 1.  $\lambda$  is a scalar weight.

#### 4.4 Sequential Model

Both the pipeline model and the joint model are redundant since they require separate modules to generate QAPs. And errors introduced by the pipeline or separate training could result in inconsistency. Therefore, we propose a sequential model with just one decoder as shown in Figure 2. The sequential model concatenates question and answer as an integral sentence and inserts a reserved token “<sep>” between them. This sentence is denoted as  $T$ . Similar to the pipeline model, the order of question and answer is also changeable.

The baseline sequential model aims to maximize directly the log likelihood represented as  $P_\Omega(T|I)$ . In the variation version, the prior  $P_\Phi(Z|I)$  is alike with  $P_{\Phi_X}(Z_X|I)$  in Equation 2, and the posterior  $P_\Psi(Z|I, T)$  is alike with  $P_{\Psi_Y}(Z_Y|I, X, Y)$  in Equation 3.

#### 4.5 Region Representation Scaling

We initially use the latent variable  $Z$  to transform the state of the decoder. To make full use of  $Z$ , we also propose a novel strategy to scale the region representation. The core idea is that since the target QAP is tightly related to only a few regions in the image and its information is included in the latent variable, we can scale the region representations to highlight those related and weaken those unrelated before decoding. Specifically, we assign a weight to the representations of each region:

$$\begin{aligned} h_i &= w_i h_i \\ w_i &= \min \left[ 1, R \times \text{Softmax}(\mathbf{M}(Z)^\top h_i) \right]. \end{aligned} \quad (5)$$

Then we use the scaled representation for the subsequent decoding. Note that this scaling mechanism can be used in all variational models.

#### 4.6 Consistency Evaluator

Consistency evaluation is critical in our work. However, there is no existing automatic metrics. Inspired by the work in semantic evaluation (Wieting et al., 2019), we devise an evaluation model to measure the consistency of generated QAPs with given image:

$$s = \text{Sigmoid} \circ \mathbf{M} \circ f(h_Q \oplus h_A, \mathbf{h}_I). \quad (6)$$

It will return a scalar score  $s$  between [0,1]. The score will be high if the QAP is consistent with the image. Here,  $h_Q$  and  $h_A$  are the representation of question and answer, respectively.

Training this evaluator requires both positive and negative samples. We take all original samples in the dataset as consistent, i.e., positive. We build the negative samples dynamically for each mini-batch, which contains images, questions, and answers. Specifically, the negative samples are generated via selecting one of the following four actions randomly and applying it to those positive in the mini-batch: 1) shuffling images, 2) shuffling questions, 3) shuffling answers, and 4) shuffling all of them. Then we feed the model with both the positive and the generated negative. The evaluator is trained with mean square error loss.

### 5 Experiment Setup

In this section, we give a description of datasets and evaluation metrics. Other settings including implementation details are in the Appendix.

#### 5.1 Dataset

We conduct experiments on two visual question answering datasets, VQA2.0 (Goyal et al., 2017) and Visual-7w (Zhu et al., 2016). In these two datasets, each image could map to multiple target QAPs. Because the official test set of VQA2.0 is not public, we use the official development set as the test set in our paper and randomly select ten thousand samples from the train set as our development set. For the Visual-7w, we take no extra operations. Table 2 shows the statistics of these two datasets<sup>2</sup>.

<sup>2</sup>We use **K** to represent thousand.

Dataset	Images (K)	QAPs (K)
VQA2.0	10.0/6.1/4.8	70.6/10.0/36.9
Visual-7w	14.4/5.7/8.6	69.8/28.0/42.0

Table 2: The statistics of two datasets (Train/Dev/Test).

## 5.2 Evaluation Metrics

In the following experiments, we evaluate our models with both automatic metrics and manual inspection. We mainly focus on the diversity and consistency of generated results. On the diversity side, we use Distinct (Li et al., 2016), a common metric for diversity, to measure the ratio of unique n-grams in the text. We adopt Distinct-4 (denoted as **D**) to evaluate the generation by concatenating the question and answer together. We also report the number of unique QAPs of the generation result (denoted as **N**). On the consistency side, we use our consistency evaluation model (Section 4.6) to indicate whether the QAP is consistent with the image. If the output score is greater than 0.5, we consider the result is consistent, otherwise inconsistent. We report two metrics for consistency, the percentage of consistent QAPs (denoted as **P**), and the average score of generation result (denoted as **S**).

We also evaluate the diversity and the consistency manually on the VQA2.0 dataset. Specifically, four human annotators perform diversity and consistency evaluation on randomly selected images. Each image could contain three to ten generated QAPs. We ask every human annotator to rate the QAPs in terms of the above two metrics. The evaluation result will be transformed into a score of [0,1] (higher score means better performance). Detailed guidelines for different ratings are provided to the human judges (see Appendix).

## 5.3 Implementation Details

In all experiments, we use pre-trained ResNet-50 as the CNN encoder. Both the sequence encoder and decoder are long short-term memory networks with two layers. We do not tune the hyperparameters elaborately towards the dataset. Therefore, all models on the two datasets share the same parameter settings. All the representations and latent variables are 512-dimensional vectors. The question and answer share the same dictionary that keep all tokens. The word embedding is initialized using Glove. We dropout all models with a ratio of 0.1. To avoid posterior collapse, we use free-bits of 5

to the KL term in the ELBO. We also use free-bits of 0.03 to the KL term in the attention alignment loss, to allow existence of divergence between the two attentions. We set the weight  $\lambda$  of  $\mathcal{L}_{attn}$  to 0.5. We train all the model 40 epochs with batch size 256 on two Nvidia Titan RTXs. The parameters are updated by Adam optimizer, with the initial learning rate  $1e-3$ . The learning rate decays with a ratio of 0.5 if the model has not improved for five consecutive epochs on the development set.

The VQA model and the VQG model in the paper share the same encoder-decoder architecture. More specifically, they share modules with the *Y*-model in the baseline pipeline model. They take the image, the answer (for VQG) or the question (for VQA) as input for generation. The hyperparameters for the baseline VQA, the baseline VQG, and pre-training remain consistent with those mentioned above. Except for the initial learning rate and training epochs, other hyperparameters keep unchanged in the fine-tuning process.

Finally, all models are implemented using the framework Fairseq<sup>3</sup>, and the source code is available at <https://github.com/LtECoD/vqapg>.

Dataset	Train(%)	Dev(%)	Test(%)
VQA2.0	89.6	87.9	89.2
Visual-7w	87.2	85.3	89.6

Table 3: Accuracy of the consistency evaluator. The train and development sets contain negative samples, while the test set contains only positive samples.

## 6 Results

### 6.1 Performance of Consistency Evaluator

We present the performance of our consistency evaluator in Table 3. We can find the evaluator performs well to distinguish consistent and inconsistent samples, as the accuracy on the training set and development set exceeds 85% already. Even on the test set, which contains only consistent samples, the accuracy can approach 90%. It indicates the evaluator is competent to measure the consistency.

### 6.2 Quantitative Analysis

The performance of our three models on the two datasets is shown in Table 4. Here we only present the pipeline model and the sequential model that

<sup>3</sup><https://github.com/pytorch/fairseq>.

Model	Diversity		Consistency	
	D(%)	N(K)	P(%)	S
VQA2.0				
Pipeline	2.77	1.89	98.68	0.88
Pipeline*	<b>27.30</b>	<b>27.90</b>	71.38	0.64
Joint	2.39	2.36	46.50	0.43
Joint*	11.47	15.10	64.02	0.58
Sequential	1.96	1.33	<b>98.95</b>	<b>0.89</b>
Sequential*	6.79	6.71	96.02	0.86
Visual-7W				
Pipeline	3.29	2.68	98.20	0.83
Pipeline*	<b>20.01</b>	<b>24.21</b>	73.50	0.63
Joint	4.15	4.58	23.61	0.21
Joint*	12.15	15.14	79.34	0.66
Sequential	2.10	1.73	<b>98.54</b>	<b>0.84</b>
Sequential*	7.49	8.17	97.44	0.81

Table 4: Automatic evaluation results on the two datasets. The superscript asterisk indicates the variational version.

generate answers first. On both datasets, the variational pipeline model obtains the best diversity. By referring to the Table 2, we can find there are about 75% (on VQA2.0) and 50% (on Visual-7w) generated QAPs are unique. The baseline sequential model achieves the best consistency, which even reaches 98.95. Compared with these two models, the joint model is much inferior to both diversity and consistency.

On the other hand, by comparing the baseline and the variational models. It is obvious that the benefit of variational to diversity is significant for the pipeline and joint models. For example, the variational pipeline’s Distinct is almost ten times the baseline on the VQA2.0. However, the diversity gain is not evident for the sequential model. As for the consistency, both the pipeline model and the sequential model decreases, especially for the pipeline model. However, for the joint model, the consistency even raises from 23.61% to 79.34% on the visual-7w, achieving a considerable improvement and even surpassing the pipeline model.

These results prove that: 1) Although diversity is largely improved for the pipeline model, noises will also be introduced through latent variables, thereby damaging consistency; 2) The latent variable contains information of the target QAP and reduces the target space so that the consistency is largely

improved for the joint model; 3) The baseline sequential model’s diversity is improved slightly by variation, indicating it is not as sensitive to the latent variable as the former two models.

Model	Diversity		Consistency	
	D(%)	N(K)	P(%)	S
VQA2.0				
Pipeline*	<b>27.30</b>	<b>27.90</b>	71.38	<b>0.64</b>
-scaling	26.47	26.95	<b>73.91</b>	0.60
Joint*	11.47	15.10	<b>64.02</b>	<b>0.58</b>
-scaling	<b>12.30</b>	<b>15.92</b>	55.39	0.51
$-\mathcal{L}_{attn}$	11.39	15.14	57.57	0.53
Sequential*	<b>6.79</b>	<b>6.71</b>	<b>96.02</b>	<b>0.86</b>
-scaling	5.7	5.56	94.65	0.84
Visual-7w				
Pipeline*	20.01	<b>24.21</b>	<b>73.50</b>	<b>0.63</b>
-scaling	<b>20.89</b>	23.45	71.29	0.61
Joint*	12.15	15.14	<b>79.34</b>	<b>0.66</b>
-scaling	<b>13.70</b>	<b>17.64</b>	72.66	0.62
$-\mathcal{L}_{attn}$	11.20	14.58	76.03	0.64
Sequential*	7.49	8.17	<b>97.44</b>	<b>0.81</b>
-scaling	<b>8.97</b>	<b>8.64</b>	96.75	0.80

Table 5: Ablation study on VQA2.0 and Visual-7w.

### 6.3 Ablation Study

To verify the effectiveness of our proposed methods, including the region representation scaling and attention alignment. We study their impact on the three models. Table 5 shows the results. As observed, the scaling mechanism shows a benefit to the consistency for all models. Especially for the joint model on VQA2.0, the consistent percentage raise from 55.39 to 64.02, achieving a 8.63 improvement. However, it could also reduce diversity as the diversity metrics will suffer a little drop in most cases. It indicates there is a trade-off between diversity and consistency.

On the other hand, removing attention alignment loss  $\mathcal{L}_{attn}$  leads to significantly lower consistency for the joint model. Besides, the diversity is also reduced slightly. It indicates that  $\mathcal{L}_{attn}$  contributes to improving consistency without sacrificing diversity.

We also investigate the impact of generation order on diversity and consistency for the pipeline model and the sequential model. We compare the

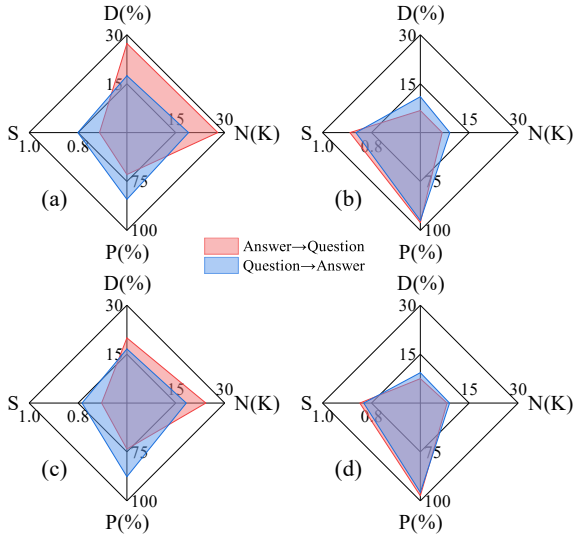


Figure 5: The performance of the variational pipeline model and the variational sequential model with different generation orders on the two datasets. (a): The pipeline model on VQA2.0; (b): The sequential model on VQA2.0; (c): The pipeline model on Visual-7w; (d): The sequential model on Visual-7w.

models with different generations orders on the two datasets. Figure 5 shows the results. The impact of generation order on the pipeline model is evident. It shows the pipeline model that generates answers first tends to obtain a good diversity while the model generates questions first could obtain a better consistency. Compared to the pipeline model, the impact of generation order on the sequential model is slighter because the volatility caused by the generation order is small. However, it shows a contrary phenomenon that the sequential model generating question first obtains a better diversity and the sequential model generating answer first obtains a better consistency. Such evidence indicates the trade-off between diversity and consistency again.

## 6.4 Human Evaluation

To better evaluate the quality of the generated QAPs, we conduct diversity and consistency evaluation manually. We evaluate three variational models. We randomly select 50 shared images from the test set of VQA2.0 and give their generated QAPs to annotators. For each image, we remove extra repeated QAPs. Table 6 shows the results.

Both the pipeline and the joint models generate more than 350 QAPs, while the sequential model generates 100 fewer samples, indicating the inferiority for generating diverse QAPs. The pipeline

Model	Count	Diversity	Consistency
Pipeline*	<b>365</b>	<b>0.75</b>	0.58
Joint*	355	0.62	0.64
Sequential*	248	0.46	<b>0.73</b>

Table 6: Human evaluation results towards that diversity and consistency on the VQA2.0.

model obtains the highest diversity score, and the sequential obtains the lowest, indicating the same result as the quantitative evaluation. However, for the consistency score, the joint model shows an obvious advantage over the pipeline model. Nevertheless, the sequential model only achieves a consistency score of 0.73, not well-matched with the result in Table 4 which indicates more than 97% generated QAPs are consistent. We argue that the pipeline and the sequential models are good at capturing the linguistic features of QAPs (such as co-occurrence) because there is information flow between the question and answer. However, the joint model is blind to such information and can only capture them through latent variables. Consequently, the joint model obtains lower automatic consistency metrics but a high human evaluation score. It also indicates the benefit of the latent variable to improve the consistency of the joint model.

## 6.5 Case Study

We present several generated QAPs given the second image in Figure 1. The examples are shown in Table 7. As observed, all three models can detect the objects in the image. The sequential model

	Questions	Answers
Pipeline*	What room is in the picture?	Office
	Where is the picture taken?	Office
	<i>What color is on the computer screen?</i>	<i>Orange</i>
	<i>What is on top of the computer?</i>	<i>Wire</i>
Joint*	What kind of computer is this?	Laptop
	What is on the screen?	Windows
	<b>What brand is the computer?</b>	<b>Apple</b>
	<i>What color is the computer?</i>	<i>Unknown</i>
Sequential*	Where is the printer?	Nowhere
	What color is the keyboard?	Black
	<b>What brand is the computer?</b>	<b>Dell</b>
	What room is this?	Office

Table 7: Cases of generated QAPs given the image in Figure 1 by different models. Italics means inconsistent QAP, and bold indicates it is difficult to judge the consistency by human.



Model	BLEU	METEOR	ROUGE
VQG			
Baseline	25.37	29.74	62.45
+Pipeline*	<b>26.42</b>	<b>30.54</b>	<b>63.29</b>
+Joint*	25.24	29.90	62.46
+Sequential*	26.04	30.41	62.98
VQA			
Baseline	16.26	22.25	57.85
+Pipeline*	<b>17.10</b>	22.28	57.96
+Joint*	17.05	22.34	57.86
+Sequential*	17.09	<b>22.63</b>	<b>58.27</b>

Table 8: Performance of two baseline models and their three fine-tuned versions pre-trained on the different generated corpus.

even asks nonexistent object “*printer*”, but it also gives correct answer “*nowhere*”. However, it also reveals some problems. Although they can generate rational questions, the joint model and the pipeline model could give wrong answers. The sequential model shows no apparent errors, but it still produces QAPs that humans cannot judge. It indicates that the models could be heavily dependent on the linguistic features and ignore the association with the image.

## 6.6 Effect to VQG and VQA

To illustrate the effectiveness of VQAPG to downstream applications, we use three variational models to generate five times training QAPs to pre-train a VQG model and a VQA model. The implementation of the two models is shown in the Appendix. Specifically, we pre-train the two models on the generated corpus and fine-tune them on the original training set. The number of epochs is 20, and the initial learning rate is  $3e-4$  for fine-tuning, while other hyper-parameters remain unchanged (see Appendix). We use BLEU-4, METEOR, and ROUGE-L as evaluation metrics<sup>4</sup>. Table 8 shows the performance of the baseline VQG and VQA model and their three fine-tuned versions. We can observe that pre-training on the corpus generated by the variational pipeline model and the sequential model can improve VQG and VQA more significantly than the joint model. Overall, such evidence indicates the benefit of the VQAPG task to the VQG and

<sup>4</sup>These metrics are common in VQG. Since answers in Visual-7W typically contains several tokens, we still use these metrics in VQA.

VQA. Another observation is that the variational pipeline model and the variational sequential model contribute differently to VQG and VQA. By referring to the analysis mentioned above of diversity and consistency, we can conclude that VQG focuses more on diversity, while VQA focuses more on consistency.

## 7 Conclusion

In this paper, we propose a novel task, VQAPG, which generates question-answer pairs from images. We also propose three models to perform this task. Targeting on diversity and consistency, we integrate variational inference to these models and propose a series of actions, including region representation scaling and attention alignment. To evaluate the consistency automatically, we devise an evaluator. We evaluate our models on two datasets: VQA2.0 and Visual-7w. The results show each model has its own merits. Overall, they perform well to generate diverse and consistent QAPs.

On the other hand, there are still limitations in our works. For example, there is a trade-off between diversity and consistency; the generated question is typically one-hop, requiring no extra reasoning; the latent variable is uncontrollable and could introduce unexpected linguistic features to the decoder, bringing inconsistent QAPs; the consistency evaluator needs more robust training strategy. In future works, we will explore generating consistent deep question-answer pairs.

## References

- Satanjeev Banerjee and Alon Lavie. 2005. **METEOR: An automatic metric for MT evaluation with improved correlation with human judgments**. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, and D. Batra. 2017. **Visual dialog**. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1080–1089.
- Xinya Du, Junru Shao, and Claire Cardie. 2017. **Learning to ask: Neural question generation for reading comprehension**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.

- Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. 2017. [Making the v in vqa matter: Elevating the role of image understanding in visual question answering](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334.
- Michael Heilman and Noah A. Smith. 2010. [Good question! statistical ranking for question generation](#). In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617, Los Angeles, California. Association for Computational Linguistics.
- U. Jain, Z. Zhang, and A. Schwing. 2017. [Creativity: Generating diverse questions using variational autoencoders](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5415–5424.
- Yanghoon Kim, Hwanhee Lee, Joongbo Shin, and Kyomin Jung. 2019. [Improving neural question generation using answer separation](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6602–6609.
- Diederik P. Kingma and Max Welling. 2014. [Auto-encoding variational bayes](#). In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.
- Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2019. [Information maximizing visual question generation](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2008–2018.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. [Visual genome: Connecting language and vision using crowdsourced dense image annotations](#). *International journal of computer vision*, 123(1):32–73.
- Igor Labutov, Sumit Basu, and Lucy Vanderwende. 2015. [Deep questions without deep understanding](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 889–898, Beijing, China. Association for Computational Linguistics.
- Dong Bok Lee, Seanie Lee, Woo Tae Jeong, Donghwan Kim, and Sung Ju Hwang. 2020. [Generating diverse and consistent QA pairs from contexts with information-maximizing hierarchical conditional VAEs](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 208–224, Online. Association for Computational Linguistics.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. [A diversity-promoting objective function for neural conversation models](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.
- Y. Li, N. Duan, B. Zhou, X. Chu, W. Ouyang, X. Wang, and M. Zhou. 2018. [Visual question generation as dual task of visual question answering](#). In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6116–6124.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. [Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.
- Xiyao Ma, Qile Zhu, Yanlin Zhou, and Xiaolin Li. 2020. [Improving question generation with sentence-level semantic matching and answer position inferring](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8464–8471.
- Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. [Generating natural questions about an image](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1802–1813, Berlin, Germany. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Badri Narayana Patro, Sandeep Kumar, Vinod Kumar Kurmi, and Vinay Namboodiri. 2018. [Multimodal differential network for visual question generation](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4002–4012, Brussels, Belgium. Association for Computational Linguistics.
- Yi Ren, Jinglin Liu, Xu Tan, Zhou Zhao, Sheng Zhao, and Tie-Yan Liu. 2020. [A study of non-autoregressive model for sequence generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 149–159, Online. Association for Computational Linguistics.
- Hung-Ting Su, Chen-Hsi Chang, Po-Wei Shen, Yu-Siang Wang, Ya-Liang Chang, Yu-Cheng Chang, Pu-Jen Cheng, and Winston H. Hsu. 2021. [End-to-End Video Question-Answer Generation with Generator-Pretester Network](#). *arXiv:2101.01447*.

- Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. [VI-bert: Pre-training of generic visual-linguistic representations](#). In *Eighth International Conference on Learning Representations (ICLR)*.
- Sandeep Subramanian, Tong Wang, Xingdi Yuan, Saizheng Zhang, Adam Trischler, and Yoshua Bengio. 2018. [Neural models for key phrase extraction and question generation](#). In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 78–88, Melbourne, Australia. Association for Computational Linguistics.
- Xingwu Sun, Jing Liu, Yajuan Lyu, Wei He, Yanjun Ma, and Shi Wang. 2018. [Answer-focused and position-aware neural question generation](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3930–3939, Brussels, Belgium. Association for Computational Linguistics.
- Y. Sun, D. Tang, N. Duan, T. Qin, S. Liu, Z. Yan, M. Zhou, Y. Lv, W. Yin, X. Feng, B. Qin, and T. Liu. 2020. [Joint learning of question answering and question generation](#). *IEEE Transactions on Knowledge and Data Engineering*, 32(5):971–982.
- Siyuan Wang, Zhongyu Wei, Zhihao Fan, Yang Liu, and Xuanjing Huang. 2019. [A multi-agent communication framework for question-worthy phrase extraction and question generation](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7168–7175.
- Zichao Wang, Andrew Lan, Weili Nie, Andrew Waters, Phillip Grimaldi, and Richard Baraniuk. 2018. [Qg-net: a data-driven question generation model for educational content](#). In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, pages 1–10.
- John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. [Beyond BLEU: training neural machine translation with semantic similarity](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.
- Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Hengel. 2016. [Visual question answering: A survey of methods and datasets](#). *Computer Vision and Image Understanding*, 163.
- Shijie Zhang, Lizhen Qu, Shaodi You, Zhenglu Yang, and Jiawan Zhang. 2017. [Automatic generation of grounded visual questions](#). In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4235–4243.
- Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. [Unified vision-language pre-training for image captioning and vqa](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13041–13049.
- Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2018. [Neural question generation from text: A preliminary study](#). In *Natural Language Processing and Chinese Computing*, pages 662–671, Cham. Springer International Publishing.
- Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. 2016. [Visual7w: Grounded question answering in images](#). In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4995–5004.

## A Autoregressive Generation

We adopt an autoregressive generation for sequences in this paper. Therefore, all the likelihood can be expanded further. For example:

$$P_{\Omega_Q}(Q|I) = \prod_{i \leq m} P_{\Omega_Q}(q_i|I, q_{<i})$$

$$P_{\Omega_Q}(Q|I, Z) = \prod_{i \leq m} P_{\Omega_Q}(q_i|I, Z, q_{<i}),$$

## B Derivation of ELBO

### B.1 The Pipeline Model

The pipeline model contains two sub-models, so there are two ELBOs.

Firstly, for the  $X$ -model, the Kullback-Leibler (KL) divergence between the estimated posterior distribution and the true posterior distribution of  $Z_X$  is:

$$\text{KL}[P_{\Psi_X}(Z_X|I, X)||P(Z_X|I, X)].$$

It can be expanded as:

$$\begin{aligned} & \text{KL}[P_{\Psi_X}(Z_X|I, X)||P(Z_X|I, X)] \\ &= \mathbb{E}_{Z_X \sim P_{\Psi_X}} \left[ \log \frac{P_{\Psi_X}(Z_X|I, X)}{P(Z_X|I, X)} \right] \\ &= \mathbb{E}_{Z_X \sim P_{\Psi_X}} \left[ \log \frac{P_{\Psi_X}(Z_X|I, X)P(X|I)}{P(X|I, Z_X)P(Z_X|I)} \right] \\ &= \text{KL}[P_{\Psi_X}(Z_X|I, X)||P(Z_X|I)] \\ & \quad + \log P(X|I) \\ & \quad - \mathbb{E}_{Z_X \sim P_{\Psi_X}} [\log P(X|I, Z_X)]. \end{aligned}$$

Then:

$$\begin{aligned} & \log P(X|I) \geq \\ & \log P(X|I) - \text{KL}[P_{\Psi_X}(Z_X|I, X)||P(Z_X|I, X)] \\ &= \mathbb{E}_{Z_X \sim P_{\Psi_X}} [\log P(X|I, Z_X)] \\ & \quad - \text{KL}[P_{\Psi_X}(Z_X|I, X)||P(Z_X|I)]. \end{aligned}$$

Because we use  $P_{\Phi_X}(Z_X|I)$  to estimate the true prior  $P(Z_X|I)$ , and use  $P_{\Omega_X}(X|I, Z_X)$  to estimate the likelihood  $P(X|I, Z_X)$ . We can get the ELBO:

$$\begin{aligned} \log P(X|I) \geq & \mathbb{E}_{Z_X \sim P_{\Psi_X}} [\log P_{\Omega_X}(X|I, Z_X)] \\ & - \text{KL} [P_{\Psi_X}(Z_X|I, X) || P_{\Phi_X}(Z_X|I)]. \end{aligned}$$

Similarly, for the  $Y$ -model, the KL divergence between estimated posterior and true posterior of  $Z_Y$  is:

$$\text{KL} [P_{\Psi_Y}(Z_Y|I, X, Y) || P(Z_Y|I, X, Y)].$$

Through the sample derivation process of ELBO for  $X$ -model, we can get the ELBO of  $Y$ -model is:

$$\begin{aligned} \log P(Y|I, X) \geq & \mathbb{E}_{Z_Y \sim P_{\Psi_Y}} [\log P_{\Omega_Y}(Y|I, X, Z_Y)] \\ & - \text{KL} [P_{\Psi_Y}(Z_Y|I, X, Y) || P_{\Phi_Y}(Z_Y|I, X)]. \end{aligned}$$

## B.2 The Joint Model

We also first present the KL divergence between the estimated posterior and the true posterior of  $Z$ :

$$\text{KL} [P_{\Psi}(Z|I, Q, A) || P(Z|I, Q, A)].$$

Then the KL divergence is expanded as:

$$\begin{aligned} & \text{KL} [P_{\Psi}(Z|I, Q, A) || P(Z|I, Q, A)] \\ &= \mathbb{E}_{Z \sim P_{\Psi}} \left[ \log \frac{P_{\Psi}(Z|I, Q, A)}{P(Z|I, Q, A)} \right] \\ &= \mathbb{E}_{Z \sim P_{\Psi}} \left[ \log \frac{P_{\Psi}(Z|I, Q, A)P(Q, A|I)}{P(Q, A|I, Z)P(Z|I)} \right] \\ &= \text{KL} [P_{\Psi}(Z|I, Q, A) || P(Z|I)] + \log P(Q, A|I) \\ & \quad - \mathbb{E}_{Z \sim P_{\Psi}} [\log P(Q, A|I, Z)]. \end{aligned}$$

We estimate the true prior  $P(Z|I)$  with  $P_{\Phi}(Z|I)$ . And the joint model assumes  $P(Q, A|I) = P(Q|I)P(A|I)$ , so  $P(Q, A|I)$  is estimated as  $P_{\Omega_Q}(Q|I)P_{\Omega_A}(A|I)$ . Then the ELBO for the joint model is:

$$\begin{aligned} \log P(Q, A|I) \geq & \mathbb{E}_{Z \sim P_{\Psi}} [\log P_{\Omega_Q}(Q|I, Z)P_{\Omega_A}(A|I, Z)] \\ & - \text{KL} [P_{\Psi}(Z|I, Q, A) || P_{\Phi}(Z|I)]. \end{aligned}$$

## B.3 The Sequential Model

As the sequential model concatenates question and answer into an integral sequence, the derivation of ELBO is same as the  $X$ -model in the pipeline model.

## C Guidelines of Human Evaluation

### C.1 Diversity

We ask human annotators to inspect the diversity of a group QAPs generated from a common image and score them from two aspects: the question type and the objects that appeared. The annotator computes the percentage of unique question types and unique objects. We then average them as the diversity score for the QAP group. Taking the results of the joint model in Table 7 of the paper as an example. This group of QAPs includes four question types: “*what kind*”, “*what is*”, “*what brand*”, and “*what color*”. The appeared objects are “*computer*”, and “*screen*”. Therefore, the diversity score for this group is  $(4/4+2/4)/2 = 0.75$ . The overall score of all samples is the average of all groups. We also take the number of QAPs into consideration for diversity score. Specifically, we normalize the model’s final diversity score with a ratio, which is computed by dividing the count of model’s generated QAP by the maximum count of all models. The final diversity score is the average of all annotators.

Model	BLEU	METEOR	ROUGE
VQA2.0			
Pipeline	39.66	27.99	65.88
Pipeline*	16.69	19.46	49.83
Joint	<b>45.18</b>	28.56	<b>67.81</b>
Joint*	33.93	23.62	61.02
Sequential	42.40	<b>28.69</b>	67.29
Sequential*	33.57	24.45	60.05
Visual-7W			
Pipeline	19.14	21.84	56.36
Pipeline*	11.23	19.16	50.13
Joint	19.83	21.44	55.34
Joint*	17.29	20.78	54.19
Sequential	<b>20.21</b>	<b>21.94</b>	<b>56.73</b>
Sequential*	17.19	21.20	53.79

Table 9: Supplemental automatic evaluation results on the two datasets.

### C.2 Consistency

Human annotators evaluate the consistency for each QAP from two aspects as well. One is that is the question answerable for the given image. Another is that is the answer correct. The scoring criteria is: **0**-the question is not answerable. **1**-the

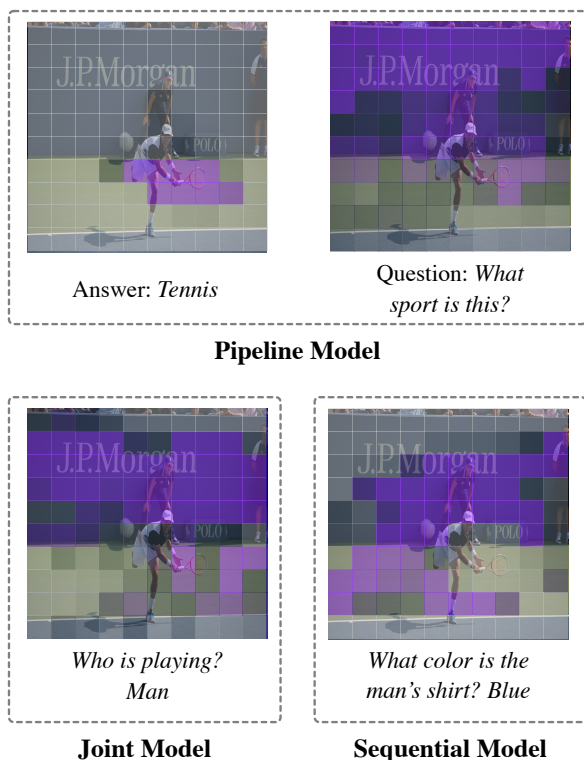


Figure 6: Weights for regions of different models. Deeper color means bigger weight.

question is answerable, but the answer is incorrect; **2**-the question is answerable, and the answer is correct. For those answers that human annotators cannot judge the correctness, we take them as consistent. Taking output of the joint model in Table 7 of the paper as an example. The score of “*What is on the screen? Windows*” and “*What brand is the computer? Apple*” is 2. While the score of “*What color is the computer? Unknown*” is 1. Then this score is divided by 2 as the final consistency score. Same as the diversity evaluation, the final consistency score is the average of all annotators.

## D More Results

### D.1 Automatic Evaluation Results

In addition to metrics of diversity and consistency introduced in the paper, we also measure more n-gram based metrics, including BLEU-4, METEOR and ROUGE-L. Table 9 shows the result. As we can see, all the baseline models achieve higher n-gram scores than the variational models. Among the three models, the pipeline model drops most from the baseline to the variational. On the visual-7w dataset, the baseline sequential model wins all metrics. However, on the VQA2.0 dataset, it is the baseline joint model that wins the BLEU and

ROUGE. By comparing with the paper results, we can find that better BLEU, METEOR, or ROUGE does not mean better diversity and consistency. Therefore, such n-gram based metrics are insufficient to measure the degree of diversity and consistency, although they reflect the overlapping between generated results and gold references.

### D.2 Region Representation Scaling

To inspect the effectiveness of our proposed region representation scaling mechanism. We randomly select an image and visualize region weights  $w$  of every model. The results are shown in Figure 6.

The pipeline model generates the answer first. As we can see, the answer-model allocates bigger weights to a few regions and generates answer “*tennis*”, indicating it detects the answer information from the latent variable successfully. While the question-model allocates weights more broadly and produces the question “*What sport is this?*”, indicating question generation requires focus more regions of the image.

Both the joint model and sequential model illustrate a similar phenomenon with the question-model of the pipeline, i.e., they require more regions for generation. Although the joint model

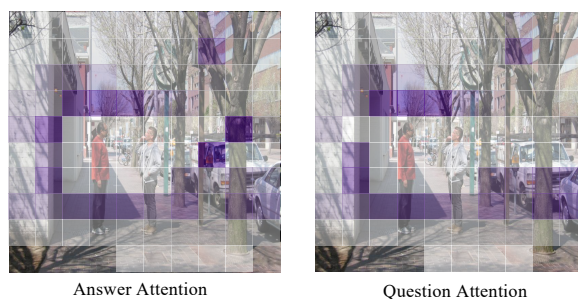


Figure 7: The average attentions of the answer decoder (left) and the question decoder (left) for the QAP “*Where was this photo taken? In the city.*”. Note that darker colors mean bigger weights.

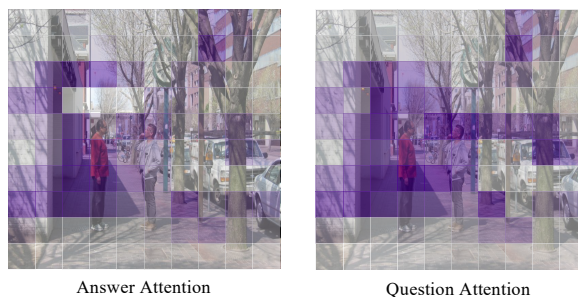


Figure 8: The average attentions for the QAP “*How many people are there? One.*”

generates a consistent QAP, the visualized weights indicate the answer “man” refers to the man behind. The sequential model is good proof. By analyzing its generated QAP, we can easily find the answer “Blue” refers to the man behind, instead of the man in front. All these evidence imply the three models can mine the QAP information from latent variables.

### D.3 Attention Alignment

To inspect our proposed attention alignment mechanism for the joint model, we also visualize the average attentions of the question decoder and the answer decoder by randomly selecting an example as shown in Figure 7 and Figure 8. We sample two QAPs generated by the joint model. One is “Where was this photo taken? In the city.”, and another is “How many people are there? One.”. Obviously, the first is consistent, while the second is not because the right answer should be “two”.

Figure 7 shows two average attentions for the consistent QAP. As we can see, the two attentions are very close. It indicates the attention alignment mechanism works successfully for this QAP generation.

On the other hand, Figure 8 shows two attentions for another inconsistent QAP. We can easily find the divergence between these two attentions. More interestingly, the answer is “one” and its attention only cover the left people. But the correct answer is “two,” and the question attention covers two peoples in the image correspondingly.

## E Model Size

We report parameter numbers of all models in Table 10. As observed, the size decreases from the pipeline model to the joint model. The variational pipeline is 50M bigger than the variational sequential model. Since the capacity of small models is limited, the model size provides a possible interpretation for the disadvantage of the sequential model in diversity.

Model	Size(M)	Model	Size(M)
Pipeline	61.15	Pipeline*	90.30
Joint	36.08	Joint*	65.82
Sequential	24.46	Sequential*	41.94

Table 10: Number of parameters of each model.