C3VQG: Category Consistent Cyclic Visual Question Generation

Shagun Uppal¹*, Anish Madan¹*, Sarthak Bhagat¹*
Yi Yu², Rajiv Ratn Shah¹
¹IIIT-Delhi, India; ²NII, Japan
{shagun16088,anish16223,sarthak16189,rajivratn}@iiitd.ac.in,yiyu@nii.ac.jp

ABSTRACT
Visual Question Generation (VQG) is the task of generating natural questions based on an image. Popular methods in the past have explored image-to-sequence architectures trained with maximum likelihood which have demonstrated meaningful generated questions given an image and its associated ground-truth answer. VQG becomes more challenging if the image contains rich context information describing its different semantic categories. In this paper, we try to exploit the different visual cues and concepts in an image to generate questions using a variational autoencoder (VAE) without ground-truth answers. Our approach solves two major shortcomings of existing VQG systems: (i) minimize the level of supervision and (ii) replace generic questions with category relevant generations. Most importantly, through eliminating expensive answer annotations, the required supervision is weakened. Using different categories enables us to exploit different concepts as the inference requires only the image and category. Mutual information is maximized between the image, question, and answer category in the latent space of our VAE. A novel category consistent cyclic loss is proposed to enable the model to generate consistent predictions with respect to the answer category, reducing its redundancies and irregularities. Additionally, we also impose supplementary constraints on the latent space of our generative model to provide structure based on categories and enhance generalization by encapsulating decorrelated features within each dimension. Through extensive experiments, the proposed C3VQG outperforms the state-of-the-art visual question generation methods with weak supervision.

KEYWORDS
visual question generation, cycle consistency, multimodal

1 INTRODUCTION
Visual understanding by intelligent systems is a very interesting problem in the Computer Vision community, further accelerated by the advent of Deep Learning. Humans tend to develop different concepts about visual data depending on context, and researchers have tried to replicate this behavior in intelligent systems like conversational agents. Translating this visual understanding into language helps us evaluate the “comprehension capability” of the system. Tasks like Visual Question Answering (VQA) [1, 18, 31], Visual Question Generation (VQG) [20], and Video Captioning [5] help us benchmark it. Such tasks require us to learn multimodal representations from visual and language data. VQG is a much more open-ended and creative task than VQA in the sense that there exist many concepts in the image, and asking semantically coherent and visually relevant questions requires a system to recognize those concepts. Unlike this, in VQA the model tries to infer specific cues from the given information in order to answer the reference questions. Figure 1 illustrates some abstract concepts and the various semantics that are captured via broad categories that we considered for question generation. Each category is distinctive enough to be exclusive from others and at the same time, covers a broad range of possibilities for question generation, when an image is conditioned over it.

Modelling the task of VQG brings with it many novel conceptual discoveries about language based and visual representations but at the same time poses certain challenges: (1) There are various visual concepts in the images. (2) Questions generated need to be relevant to the image. (3) The generated question to image relation is many-to-one since multiple questions are possible for an image, and (4) Avoiding questions which invoke generic answers like “yes” or “I do not know”. For example in Figure 1, we can observe the little girl jumping, the mother trying to read something, the image is of a hotel room, there are photos hanging on top of the bed, etc. All these concepts are relevant to the image. Additionally, the questions in Figure 1 are also relevant to the image, satisfy the many-to-one criteria with respect to a number of questions for the given image, and do not invoke generic answers.

For getting human-level understanding of multimodal real-world data, system designs should be created in order to overcome such challenges. This is the reason the task of VQG has also been referred as a realization of the Visual Curiosity [28] of a system.

Previous works [14, 15, 17, 27] often use answers alongwith the image to generate relevant questions. While this approach asks questions relevant to the image (due to the answer being provided), it tends to overfit to the answer provided and does not leave room for creatively generating questions on diverse and novel concepts in the image. For the image in Figure 1, such models are likely to generate questions like “What is the girl doing?” which might be a result of the overfitting of the model on the answer “The girl is

Figure 1: An example image showing the various natural questions possible which belong to the broad categories mentioned. The categories are not too specific so as to overly constrain the network but are broad enough to encourage discovery of novel concepts.

Possible Category-Question pairs:
SPATIAL: Where are the pictures hanging?
ACTIVITY: What is the little girl doing?
BINARY: Is the lamp on?
COUNT: How many pillows are there on the bed?
COLOR: What is the color of the girl's dress?

Figure 1: An example image showing the various natural questions possible which belong to the broad categories mentioned. The categories are not too specific so as to overly constrain the network but are broad enough to encourage discovery of novel concepts.

Possible Category-Question pairs:
SPATIAL: Where are the pictures hanging?
ACTIVITY: What is the little girl doing?
BINARY: Is the lamp on?
COUNT: How many pillows are there on the bed?
COLOR: What is the color of the girl's dress?

Figure 1: An example image showing the various natural questions possible which belong to the broad categories mentioned. The categories are not too specific so as to overly constrain the network but are broad enough to encourage discovery of novel concepts.

Possible Category-Question pairs:
SPATIAL: Where are the pictures hanging?
ACTIVITY: What is the little girl doing?
BINARY: Is the lamp on?
COUNT: How many pillows are there on the bed?
COLOR: What is the color of the girl's dress?

Figure 1: An example image showing the various natural questions possible which belong to the broad categories mentioned. The categories are not too specific so as to overly constrain the network but are broad enough to encourage discovery of novel concepts.

Possible Category-Question pairs:
SPATIAL: Where are the pictures hanging?
ACTIVITY: What is the little girl doing?
BINARY: Is the lamp on?
COUNT: How many pillows are there on the bed?
COLOR: What is the color of the girl's dress?

Figure 1: An example image showing the various natural questions possible which belong to the broad categories mentioned. The categories are not too specific so as to overly constrain the network but are broad enough to encourage discovery of novel concepts.

Possible Category-Question pairs:
SPATIAL: Where are the pictures hanging?
ACTIVITY: What is the little girl doing?
BINARY: Is the lamp on?
COUNT: How many pillows are there on the bed?
COLOR: What is the color of the girl's dress?

Figure 1: An example image showing the various natural questions possible which belong to the broad categories mentioned. The categories are not too specific so as to overly constrain the network but are broad enough to encourage discovery of novel concepts.

Possible Category-Question pairs:
SPATIAL: Where are the pictures hanging?
ACTIVITY: What is the little girl doing?
BINARY: Is the lamp on?
COUNT: How many pillows are there on the bed?
COLOR: What is the color of the girl's dress?
jumping”. However, it is highly unlikely for these models to get questions like “What is the color of the girl’s dress?” or “Is the lamp on?” due to the lack of conditioning on the answer category. It restricts the many-to-one relation between an image and the possible set of questions that can be generated with respect to it. Also, this requires the dataset to be annotated with answers as well as questions which is an expensive and tedious operation.

While current works rely heavily on the availability of question-answer pairs for their method, we propose using categories instead of answers. This incorporates a weaker form of supervision, which is easy to obtain and can help in exploring various abstractions in an image as well. This also helps generate relevant questions to the image as compared to methods which simply generate questions based on an image without any constraints [9, 20, 30], thus, leading to non-diverse and often not so relevant questions. We propose a generative modelling framework, where we try to ensure the category of the generated question, so that multiple relevant category-specific question generations per image are possible.

The following are the main contributions of the paper:

- We weaken the amount of supervision on the model by removing the need of ground-truth answers during the training phase. This makes our approach smoothly generalizable and waives the requirement of availability of ground-truth answers in the dataset.
- We adopt a variational autoencoder (VAE) [12] framework to generate questions. It consists of a single combined latent space for image and category embeddings and also maximizes the mutual information between them.
- We introduce additional constraints to enforce answer category consistency by utilizing a cyclic training procedure with sequential training in two disjoint steps. This helps in ensuring that the model remains consistent with its own generations.
- We enforce center loss on the generative latent space in order to ensure clustering with respect to the answer category labels, thereby, making generations more category-specific and robust. Although, this has been employed for various biometric (image-based) applications to learn discriminative features, to the best of our knowledge our paper is the first to explore its effect in a multimodal setting.
- We also introduce a hyper-prior on learning the inverse variance of the variational latent prior to capture intrinsically independent visual features within the combined latent space. This helps us in generating more diverse questions as we observe in our results in Section 4.

Our contributions ensure that we get a diverse (see Section 4.3) and relevant (see Section 4.4) set of questions given an image and category. We evaluate our result alongside other recent approaches which do not use answers for generating questions as well as which require them.

The rest of the paper is organized as follows: In Section 2, we discuss the previous works on visual question generation and structured latent space constraints. We present our approach and details of our model in Section 3. In Section 4, we provide details about the experimental setup, evaluation metrics and discuss our qualitative and quantitative results. We present our conclusion in Section 5.

2 RELATED WORKS

In this section, we discuss relevant literature that motivates key components of the C3VQG approach. In Section 2.1, we focus on various approaches that emphasised on the task of question generation from visual inputs. This is followed by Section 2.2, where we describe appropriate studies that have remodelled their latent representations for the escalation of downstream task performance. We majorly focus on approaches that have deployed additional latent constraints for enforcing clustering or have introduced an additional hyper-prior in order to capture decorrelated features within each dimension of the latent space.

2.1 Visual Question Generation (VQG)

VQG is the task of developing visual understanding from images using cues from ground-truth answers and/or answer categories in order to generate relevant question. Various works focusing on this aspect have been deeply inspired by taking into consideration the multimodal context of natural language along with visual understanding of the input.

Mostafazadeh et al. [19] suggested relevant question as well as response generations, given an image along with the relevant conversational dialogues. With the help of the dialogues, they drew broad context about the conversation from the input image. Mostafazadeh et al. [20] focused on a different paradigm of VQG wherein the goal is to generate more engaging and high-level common sense reasoning questions about the image/event highlighted in the image. This approach shifted its focus from the objects constituting the image to the visual understanding of these systems.

Yang et al. [29] simultaneously learned VQG and VQA models to understand the semantics and entities present in the input image. The former is trained using RNNs while CNNs were used for the latter. Such an approach examines and trains the learning model on both the aspects of natural language and vision, thereby, challenging its interpretability over multimodal signals. Li et al. [15] had a similar approach of training VQA and VQG networks parallely, hence, introducing an Invertible Question-Answering network. Such a model takes advantage of the question-answer dependencies while training, then takes a question/answer as an input, in return outputting its counterpart for evaluation. Works like [24] propose a joint model for training of QA and QG task. This complements both the tasks to synchronize and to learn co-operatively but restricts their abilities to explore non-trivial aspects of generation.

Zhang et al. [30] talked about automating VQG not only with high correctness but with a high diversity in the type of questions generated. For this, they take an image and its caption as the input, as generated using a dense caption module with an LSTM-based classifier for selecting the question type. The question type along with the input image and caption and an image-caption correlation output are processed to give relevant output questions. On similar lines, Jain et al. [9] worked on generating a wide variety of questions given a single image but with generative modelling. Here, they used a VAE with a combination of LSTM networks in order to generate a diverse set of questions from a single input image.

While, prior work in VQG has spanned a wide variety of training strategies for meaningful question generation, our approach C3VQG is unique in the sense that it utilizes a mutual information
maximization technique with weak supervision. On top of it, it learns a well-structured latent space with a non-standard Gaussian prior and category-wise clustering.

2.2 Structured Latent Space Constraints

2.2.1 Center Loss for Learning Discriminative Latent Features. Center loss [25] for enforcing well-clustered latent space representations have been studied extensively in the past specifically focused on bio-metric applications [10, 25, 26]. This metric-learning training strategy works on the principle of differentiating inter-class features and penalizing the distance of embeddings from their respective class centers.

Wen et al. [26] utilized center loss for the biometric task of facial recognition. The introduction of weight sharing between softmax and the center loss reduces the computational complexity. While, the employment of an entire embedding space as the center rather than the conventionally used single point representation takes into account the intra-class variations as well. Kazemi et al. [10] also proposed a novel attribute-centered loss in order to train a Deep Coupled Convolutional Neural Network (DCCNN) for the task of sketch to photo matching using facial features.

He et al. [8] proposed a triplet-center loss that aims at further improving the differentiating power of features by not only minimizing the distance of encoding from their class centers but also by maximizing it for the class centers belonging to other classes. These discriminative latent features obtained are utilized for the task of 3D object retrieval. Ghosh and Davis [6] highlighted the impact of introduction of center loss besides the cross entropy loss in CNNs for image retrieval problems, involving very few samples belonging to each class.

Besides, the center loss when coupled with softmax loss has also been employed for emotion recognition in speech data [22].

2.2.2 Hyper-prior on Latent Spaces. Various approaches have intended to capture completely decorrelated factors of variations in the data by employing diverse training strategies like utilizing generative models to learn low-dimensional subspaces [13] or imposing a soft orthogonality constraint on the latent chunks [21]. One such effective approach is to vary the prior on the generative latent space in such a way that it intrinsically enforces independence of the captured features.

Kim et al. [11] introduced a class of hierarchical Bayesian models with certain hyper-priors on the variances of the Gaussian distribution priors in a VAE. The fact that this ensures that each captured latent feature has a different prior distribution ensures that each of them are intrinsically independent and guarantees encapsulation of admissible as well as nuisance factors simultaneously. In fact, the modified hyper-prior we apply on our latent space is an extension of the adjustable Gaussian prior model suggested in [11].

Ansari and Soh [2] also focused on capturing disentangled factors of variations in an unsupervised manner by utilizing Inverse-Wishart (IW) as the prior on the latent space of the generative model. By tweaking the IW parameter, various features in a set of diverse datasets could be captured simultaneously.

Bhagat et al. [4] utilized Gaussian processes (GP) with varying correlation structure in VAEs for the task of video sequence disentangling. The obtained latent representations were exploited for downstream tasks like video frame prediction as well.

To the best of our knowledge, center loss for latent clustering on the latent space for capturing independent factors of variation has never been deployed in a multimodal setting. We take motivation from several works that have utilized these techniques to formulate a structured latent representation in order to wield superior performance on downstream tasks.

3 PROPOSED APPROACH

We introduce C3VQG, a question generation architecture which only requires <images, questions, categories> for training, and <images, category> for inference. We propose a cyclic training approach that enforces consistency in answer categories via a two-step framework. For this, we introduce a variational autoencoder (VAE) setting which maximizes the mutual information between the question generated, image and category.

The entire training flow is illustrated in Figure 2. We divide the basic training architecture into two disjoint steps as demonstrated in Figure 2. While the first step ensures encapsulation of the image and answer-category information within the latent encoding, the second step establishes compatibility in the answer-categories predicted from the generated question with that of the ground-truth answer-categories. We formulate the latent space to contain sufficient information about the answer category besides capturing all independent features of the image in a structured manner. We do this by enforcing an additional hyper-prior on the latent space (refer Section 3.5) and including a center loss based constraint (refer Section 3.4). While the former maintains a high diversity across generated questions, the latter helps in keeping up with the relevance between the image, the answer-category and the generated question.

In this section, we begin by defining some of the notations that have been used throughout the paper. This is followed by addressing the key contrasts in building up our architecture with related recent works, and an overview of our proposed approach C3VQG. This is accompanied by describing each of its individual components alongside their motivation. Lastly, we also mention the complete training procedure and the optimization strategy used to ensure convergence.

3.1 Problem Formulation

We aim to design a generative model that encapsulates information from multimodal sources of data in the form of images and answers to generate an encoding that aids the prediction of meaningful questions.

For accomplishing this task, we have multimodal training data in the form of images and corresponding question from different answer categories. We denote all unique images by the set $I_D$, set of all unique answer categories by $C_D$, and set of all unique ground-truth questions by $Q_D$, where length of the sets are given by $n_l$, $n_c$, and $n_q$ respectively. We define our training dataset as a collection of $n$ 3-tuples, $dset = \{<i_1, q_1, C_1>, ..., <i_n, q_n, C_n>\}$. For the $k^{th}$
sample in our dataset, we have image $i_k \in I_D, q_k \in Q_D, C_k \in C_D$, as $C_D = \{C_1, C_2,...C_{nc}\}$. We denote the predicted question as $\hat{q}_{k,c}$, where $k$ denotes the sample for which the question is predicted and $C$ denotes the category ($C \in C_D$), as we generate $n_c$ questions for every sample in our training set. We also denote our latent space by $z$, and the dimensions of the combined latent space by $d$.

### 3.2 Information Maximisation VQG

We consider the case of a single image $i$, its corresponding category $C$ and the question we want to generate $q$. We define our initial model (referred as Step I in Section 3.3) by defining $p(q|i,c)$ which we get by maximizing a linear combination of mutual information $I(i,q)$ and $I(c,q)$. To avoid optimizing the gradient in discrete steps (in order to get low bias and variance of the gradient estimator), we try to learn a mapping $p_{\phi}(z|i,c)$ from the image and category to a continuous latent space which we refer to as $z$. The mapping is parameterized by $\phi$ which is learned via optimization of the following objective:

$$\max_{\phi} \quad I(q,z|i,C) + \lambda_1 I(i,z) + \lambda_2 I(C,z)$$

subject to

$$z \sim p_{\phi}(z|i,C)$$

$$q \sim p_{\phi}(q|z)$$

where $\lambda_1$ and $\lambda_2$ are the weights for the mutual information terms. The mutual information in Equation 1 is intractable as we do not know true values of the posteriors $p(z|i)$ and $p(z|C)$. So we instead try to minimize its variational lower bound (ELBO). More details on the derivation of the final objective can be found in the supplementary section. Hence, we can optimize the variational lower bound by maximizing the image and category reconstruction whilst also maximizing the ME of question generation.

### 3.3 Category Consistent Cyclic VQG (C3VQG)

We build a cyclic approach for VQG to analyze the robustness of the model in terms of its predictions and the diversity of generated questions. For this, we divide our approach into two parts. The first step homogenizes the latent representations obtained from the answer categories and the one obtained from images to form a combined latent space with a variational prior. While, the next step penalises the difference in ground-truth answer categories from the ones predicted from the generated question, enforcing congruence between them.

**Step 1: Visual Question Generation.** Using two separate encoders $g^i$ and $g^c$, we generate latent encoding $h_k^i$ and $h_k^c$ for the image $i_k$ and category label $C_k$ respectively.

$$h_k^i = g^i(i_k)$$

$$h_k^c = g^c(C_k)$$

These latent encodings are passed onto an MLP after concatenation to generate another latent representation that has a Gaussian prior associated with it. This latent representation, depicted with $z \in \mathbb{R}^d$ and forming the backbone for question generation using our approach, is given by Equation 6.

$$z_k = W_{\text{MLP}}(h_k^i \oplus h_k^c)$$

where $W_{\text{MLP}}$ depicts the weights of the MLP and $\oplus$ depicts the concatenation operator for two input vectors. The concatenation of the two encodings aids the aggregation of the information of the type of question that is supposed to be generated by the model. This latent encoding should intrinsically contain all the relevant information for the generation of the question. Therefore, it is passed through a temporal model that captures the time-varying characteristics and outputs the question related to the images on the lines of the answer category.

$$\hat{q}_{k,c} = \text{LSTM}_p(z_k)$$

Therefore, we capitalise on the ground-truth question $q_k$ for the image to impose an MLE loss on the generated question $\hat{q}_{k,c_k}$.

$$L_Q = \|q_k,c_k - \hat{q}_{k,c_k}\|_2^2$$

In order to ensure abbreviation of visual features as well as category information into the $z$-space, we pass it through two separate prediction networks, $p^i$ and $p^c$ respectively. These prediction networks are trained to reconstruct the original image and category encodings. Hence, we enforce a loss based on their predictions.

$$L_I = \|p^i(z_k) - h_k^i\|_2^2$$

$$L_C = \|p^c(z_k) - h_k^c\|_2^2$$

**Step 2: Generation Consistency Assurance.** In order to substantiate the consistency of the answer category of the generated question with the given category, we pass the generated question $\hat{q}_{k,c_k}$ through a temporal classifier $\text{LSTM}_p$ that tries to predict the answer category for the generated question.

$$c_k^{\text{pred}} = \text{LSTM}_p(\hat{q}_{k,c_k})$$

Later, we impose a cross entropy loss between the predicted and actual answer category in order to penalise any irregularities within the previous step.

$$L_{\text{cons}} = -\log c_k^{\text{pred}}$$

### 3.4 Latent Space Clustering

To ensure that our model is able to accurately predict answer categories from the latent encodings, we intend to promote well-clustered latent spaces. For this, we add structure to the latent space by imposing a constraint in the form of center loss\(^2\) [25] that aggregates the latent space into a fixed number of clusters, equal to the number of answer categories in the dataset.

The center loss helps distinguish inter-category latent features by enforcing clustering in the following way:

$$L_{\text{center}} = \|z_k - c_k\|_2^2$$

where, $c_k \in \mathbb{R}^d$ depicts the class center for all such datapoints $z_k$ (where, $k \in [1,n]$) with label $C_k$. These centers are obtained

\(^2\)https://github.com/KaiyangZhou/pytorch-center-loss
C3VQG: Category Consistent Cyclic Visual Question Generation

by averaging the features of the corresponding classes updated based on mini-batches instead of the entire training data due to computational time constraints. Additionally, the update of these centers are scaled by a constant ($<1$) to avoid sudden fluctuations.

This helps in discriminating the joint image-category representations, by casting added supervision, thereby, leading to a higher fidelity and robustness in the question generation process. The structured latent representation that is obtained as a result of applying this constraint ensures escalation of distances in the latent space between samples belonging to different classes, that in turn leads to enhanced downstream task performance.

3.5 Modified Hyper-prior on the Latent Space

We also take motivation from one of models proposed by Kim et al. [11] that introduces a modified prior on the latent space explicitly ensuring each dimension to capture completely independent features. We do this by replacing the sub-optimal Gaussian normal prior on the $z$-space by a long-tail distribution. We introduce a learnable hyper-prior on the inverse variance of the Gaussian latent prior while keeping the distribution as zero mean. We also employ a supplementary regularization term that ensures sufficient nuisance dimensions.

For this, we intend to learn the inverse variance $\alpha_j$ for each dimension $j$ of the $d$-dimensional latent space. The latent space prior can then be represented as Equation 14.

$$p(z|\alpha) = \prod_{j=1}^{d} p(z_{k,j}|\alpha_j) = \prod_{j=1}^{d} \mathcal{N}(z_{k,j}; 0, \alpha_j^{-1})$$

The modified KL-divergence and additional regularization term is of the form given by Equation 15.

$$L_{bayes} = \sum_{j=1}^{d} \mathbb{E}_{p_d(x_{cc})} \left[ KL(f(z_j|x_{cc})||\mathcal{N}(z_{k,j}; 0, \alpha^{-1}_j)) \right] + \lambda_{reg} \sum_{j=1}^{d} (\alpha_j^{-1}_j - 1)^2$$

where, $x_{cc}$ is the concatenated latent encoding of the image and category encoding, i.e., $h_i \oplus h_c$, $z$ is the latent encoding with variational prior, and $f$ is the mapping function (i.e., $f : x_{cc} \rightarrow z$). The expectation is taken over the entire probability distribution ($pd$) of $x_{cc}$.

In Equation 15, $\lambda_{reg}$ is the weight for the regularization loss that promotes sparsity and increases generalization capacity of the model.

3.6 Training Strategy and Optimization Objective

We train our model by defining a combined loss $L_{total}$ that is the weighted sum of individual loss terms. Combining Equations 8, 9, 10, 12, 13 and 15, we obtain the optimization objective as follows:

$$\min_{W} L_{total} = \min_{W} \left[ L_Q + \lambda_I L_I + \lambda_C L_C + \lambda_{cons} L_{cons} + \lambda_{center} L_{center} + \lambda_{bayes} L_{bayes} \right]$$ (16)
where, \( W \) represents the combination of all learnable parameters in the complete model and \( \lambda \) are the hyperparameters depicting the weight of each loss in the combined objective.

**Algorithm 1**: Training Algorithm for C3VQG with all components.

1. **Input**: \( \text{dset} \) containing \( n \) training tuples of form \( < i, q, C > \), multi-task loss weights for all individual losses: \( \lambda \), gradient descent learning rate \( \alpha_{LR} \).
2. **Output**: Optimal weights for all the individual components of the model \( W \).
3. Initialize \( W \) with Kaiming initialization [7].
4. for epoch \( \leftarrow 1 \) to \( \text{num\_epochs} \) do
5.   for \( k \leftarrow 1 \) to \( n \) do
6.     \( i_k, q_k, C_k \leftarrow \text{dset}[k] \)
7.     Get \( h_k^i \) and \( h_k^c \) using Equation 4 and 5.
8.     Concatenate \( h_k^i \) and \( h_k^c \) to get \( z_k \) using Equation 6.
9.     Use \( z_k \) to predict \( h_k^i \) and \( h_k^c \) using networks \( p^i \) and \( p^c \) and compute \( L_I \) and \( L_C \) using Equation 9 and 10.
10. Generate question \( q_{k,C_k} \) using Equation 7 and compute \( L_Q \) using Equation 8.
11. Predict category \( C_{k,\text{pred}} \) from generated question using Equation 11 and compute \( L_{\text{cons}} \) using 12.
12. Compute \( L_{\text{center}} \) and \( L_{\text{bayes}} \) using Equation 13 and 15 respectively.
13. Find gradient of the total loss w.r.t. \( W \), i.e. \( \nabla_W L_{\text{total}} \).
14. Take gradient descent step, \( W \leftarrow W - \alpha_{LR} \nabla_W L_{\text{total}} \).
5. end

For training our model using Algorithm 1, we use gradient descent algorithm with Adam optimizer. We train the model for 15 epochs on a machine with single GeForce GTX 1080 GPU using the PyTorch framework.

4 EVALUATION

We evaluate the performance of our approach C3VQG [3] against the state-of-the-art in VQG [9, 14, 24] using a variety of diverse quantitative metrics alongside highlighting the qualitative superiority of our approach.

4.1 Dataset Features

The VQA dataset [4] [3] consists of images along with corresponding questions and answers for each image. Krishna et al. [14] annotates the answers with a set of 15 categories and labels their top 500 answers. This makes up 82% of the entire VQA dataset consisting of 367K training and validation examples. Additional information about the entire VQA dataset is presented in the supplementary. Similar to works [9, 14, 24], we have used the validation set as our test set due to the lack of availability of ground-truth answers for the test set. We use a 80:20 training-validation split for our experiments.

4.2 Evaluation Metrics

We intend to evaluate our approach in order to compare it to the prior work in VQG using a variety of language modeling metrics including \textit{BLEU}, \textit{METEOR} and \textit{CIDEr} [25]. These metrics quantify the ability of the model to generate questions similar to the ground-truth questions for the validation set.

Additionally, we compute another quantitative metric: a variant of ROUGE [16] called as ROUGE-L. This metric quantifies the similarity between the generated and ground-truth questions by utilizing the longest common sub-sequence. The advantage of using this metric alongside others mentioned is that it takes into account any structural association present at the sentence level, thereby, capturing the longest n-gram concurrently occurring in the sequence.

We also evaluate the performance of our model against the baselines using crowd-sourced metrics for testing the relevance of the generated question with respect to the ground-truth images and answer categories. For this, we conduct a user study among 5 crowd workers in which each one is supposed to answer if the generated questions are consistent with respect to the given image and answer category.

In order to quantify the heterogeneity of generated questions, we additionally employ diversity metrics in our evaluation. For this, we compute the strength and the inventiveness. While strength is referred to as the percentage of unique generated question, inventiveness refers to the ratio of unique generated questions those were unseen during training.

4.3 Quantitative Results

In Table 1, I and II depict stage I and II respectively of our approach, \( CL \) depicts the imposed center loss on the combined latent space and \( Bayes \) represents an additional hyper-prior on the inverse variance of each latent dimension. Table 1 depicts that our approach beats the state-of-the-art performance in VQG [14] without the supervision of answers while training. The role of each component in the incremental build-up of our approach is clearly observable from the ablations reported. Additionally, it also shows the significance of cyclic consistency in answer-category for generating semantically meaningful questions. Using multiple constraints on the latent space reduces the performance slightly for Bleu-2 and Bleu-4, but we observe significant increase in other language modelling metrics. We leave certain values for ROUGE-L blank in Table 1 as some prior works [9, 24] did not employ it for their evaluation.

The reported values in Table 3 depict that our model outperforms the baselines as a result of the consistency of the generated questions and the structure present in latent space. The incorporation of the supplementary constraint on the congruence of answer category ensures that the generated question is completely relevant to the category. While, the squared \( L2 \) loss between the image encoding and the encoding generated from the combined latent space assists the relevance with respect to the image.

The superiority in the diversity of generated question by our model as depicted in Table 2 highlights that imposing a different prior on each dimension of the latent space enforces generation of a set of diversified questions from different answer categories. The performance in terms of the diversity of generated questions.
Table 1: Ablation study for different components of C3VQG using different language modeling quantitative metrics against other baselines in VQG. We compare our approach against previous works using answers as well as without answers.

| Categories          | V-IC2Q [9] | Krishna et al. [14] | C3VQG w/o Bayes | C3VQG     |
|---------------------|------------|---------------------|-----------------|-----------|
|                     | Strength   | Inventiveness       | Strength        | Inventiveness |
| count               | 15.77      | 30.91               | 26.06           | 41.30      |
| binary              | 18.15      | 41.95               | 28.85           | 54.50      |
| object              | 11.27      | 34.84               | 24.19           | 43.20      |
| color               | 4.03       | 13.03               | 17.12           | 23.65      |
| attribute           | 37.76      | 41.09               | 46.10           | 52.03      |
| materials           | 36.13      | 31.13               | 45.75           | 40.72      |
| spatial             | 61.12      | 62.54               | 70.17           | 68.18      |
| food                | 21.81      | 20.38               | 33.37           | 31.19      |
| shape               | 35.51      | 44.03               | 45.81           | 55.65      |
| location            | 34.68      | 18.11               | 45.25           | 27.22      |
| predicate           | 22.58      | 17.38               | 36.20           | 31.29      |
| time                | 25.58      | 15.51               | 34.43           | 25.30      |
| activity            | 7.45       | 13.23               | 21.32           | 26.53      |
| Overall             | 12.97      | 38.32               | 26.06           | 52.11      |

Table 2: Quantitative evaluation of C3VQG against other baselines using diversity-based metrics.
Table 3: Quantitative evaluation of C3VQG against other weakly supervised baselines using crowd-sourced metrics.

| Model               | Relevance | Image | Category |
|---------------------|-----------|-------|----------|
| V-IC2Q [9]          | 90.10     | 39.00 |          |
| Krishna et al. [14] w/o A | **98.10** | 42.70 |          |
| C3VQG w/o Bayes, CL | 98.00     | 58.40 |          |
| C3VQG               | 97.80     | **60.50** |        |

Achieved by our approach with all components beats the state-of-art in VQG even without the requirement of additional answer supervision. The difference in the strength and inventiveness values with and without the latent hyper-prior suggests that capturing decorrelated features in each latent dimension enables our model to generate non-generic questions from a divergent pool of categories.

4.4 Qualitative Results

![Qualitative results for C3VQG and Krishna et al. [14] without answers.](image)

We present a set of four generated questions (from different answer categories) for a collection of images in Figure 3. It illustrates that our approach is able to generate diverse image and category specific non-generic questions. Even for a particular category, the generated questions, although completely valid for a multiple set of images, are still not trivially replicated irrespective of the image. For example, as shown in the Figure 3, questions generated for the category “binary” are quite diverse for different images, thus, taking into consideration the context of the images as well.

Additionally in Figure 4, we also demonstrate certain cases in which the questions generated by our model belong to the specified answer categories while the baseline approach in [14] without the answer supervision fails to do so. For example, the top-left image of Figure 4, C3VQG is able to generate a question whose answer falls in the category of “color”. Whereas, for the question generated by the baseline approach [14], the answer category seems to be “object” instead of “color”.

The lack of category consistency reflected by baseline approach is well accommodated in our approach by the addition of a supplementary consistency loss. We eradicate the inconsistencies of the generated questions with the provided answer-categories by including cycle consistency in the model. This ensures that the model is confident as well as correct about its own predictions. As clearly highlighted in qualitative evaluation, questions generated by [14] make complete sense with respect to each image and are not generic questions, but it is often observed that they lack parallelism of answer-category and generated questions. This is one of the loopholes with [14] that we counter using cyclic consistency based training procedure in addition to the quantitative improvements.

5 CONCLUSION

We present a novel category-consistent cyclic training approach C3VQG for visual question generation using structured latent space. Our approach is able to generate category-specific comprehensive questions using visual features present in the image without the need of ground-truth answers. With this amount of supervision, our approach beats state-of-the-art in terms of a variety of language modeling, crowd-sourcing and diversity-based metrics. Qualitatively, our approach avoids generic question formation and is able to generate answer-category specific questions even when the former approaches fail to do so. While the cyclic training procedure aids it to generate questions consistent and relevant with the given answer category, the imposed latent structure ensures enhanced diversity of generated questions. This shows that effectively designing system configurations and imposing structured constraints can help frame better models even with minimum levels of supervision.

As a further prospect to this work, we aim to analyze the efficacy of our approach in other question generation tasks such as conversational systems. We also intend to study the effect of such constraints on other multimodal tasks like image/text retrieval, image captioning, etc. for learning robust representations.

REFERENCES

[1] Ashwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2015. VQA: Visual Question Answering. *International Journal of Computer Vision* 123 (2015), 4–31.

[2] Abdul Fatir Ansari and Harold Soh. 2018. Hyperprior Induced Unsupervised Disentanglement of Latent Representations. In *AAAI*.

[3] Stanislaw Antol, Ashwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

[4] Sarthak Bhagat, Shagun Uppal, Vivian T. Yin, and Nengli Lim. 2020. Disentangling Representations using Gaussian Processes in Variational Autoencoders for Video Prediction. *ArXiv* abs/2001.02408 (2020).

[5] Shaosiang Chen, Ting Yao, and Yu-Gang Jiang. 2019. Deep Learning for Video Captioning: A Review. In *IJCAI*.

[6] Pallab Ghosh and Larry S. Davis. 2018. Understanding Center Loss Based Network for Image Retrieval with Few Training Data. In *ECCV Workshops*.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. 2015 IEEE International Conference on Computer Vision (ICCV) (2015), 1026–1034.

[8] Xinwei He, Yang Zhou, Zhihao Zhou, Song Bai, and Xiang Bai. 2018. Triplet-Center Loss for Multi-view 3D Object Retrieval. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018), 1945–1954.

[9] Unnat Jain, Ziyu Zhang, and Alexander G. Schwing. 2017. Creativity: Generating Diverse Questions Using Variational Autoencoders. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017), 5415–5424.
10. Hadi Kazemi, Sobhan Soleymani, Ali Dabouei, Seyed Mehdi Iranmanesh, and Nasser M. Nasrabadi. 2018. Attribute-Centered Loss for Soft-Biometrics Guided Face Sketch-Photo Recognition. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2018), 612–6128.

11. Minyoung Kim, Yuting Wang, Pritish Sahu, and Vladimir Pavlovic. 2019. Bayes-Factor-VAE: Hierarchical Bayesian Deep Auto-Encoder Models for Factor Disentanglement. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019), 2979–2987.

12. Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013).

13. Jack Klys, Jake Snell, and Richard S. Zemel. 2018. Learning Latent Subspaces in Variational Autoencoders. ArXiv abs/1812.06190 (2018).

14. Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2019. Information Maximizing Visual Question Generation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019), 2008–2018.

15. Yikang Li, Nan Duan, Bolei Zhou, X. R. Chu, Wanli Ouyang, and Xiaogang Wang. 2017. Visual Question Generation as Dual Task of Visual Question Answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2017), 6116–6124.

16. Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://www.aclweb.org/anthology/W04-1013

17. Feng Liu, Tao Xiang, Timothy M Hospedales, Wanli Ouyang, and Xiaogang Wang. 2017. Visual Question Generation as Dual Task of Visual Question Answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2017), 6116–6124.

18. Mateusz Malinowski and Mario Fritz. 2014. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. arXiv cs.AI/1410.0210

19. Nasrin Mostafazadeh, Chris Brockett, William B. Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spithourakis, and Lucy Vanderwende. 2017. Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation. In IJCNLP.
In this document, we begin by deriving the variational lower bound of the objective function of step 1 of our approach. We also provide an illustration to depict the inference procedure of C3VQG. This is followed by listing the hyperparameters values and the details for VQA datasets used for training and evaluation of the C3VQG model.

1 DERIVATION
We re-iterate the objective function optimized in the main paper. The objective function is parameterized by $\phi$ which is optimized as follows:

$$\max_{\phi} I(q, z|i, C) + \lambda_1 I(i, z) + \lambda_2 I(C, z)$$

subject to:

$$z \sim p_{\phi}(z|i, C)$$

$$q \sim p_{\phi}(q|z)$$

where $\lambda_1$ and $\lambda_2$ are the weights for the mutual information terms. The mutual information in Equation 1 is intractable as we do not know true values of the posteriors $p(z|i)$ and $p(z|C)$. So we instead try to minimize its variational lower bound (ELBO). In the equations below, $\mathbb{H}$ stands for entropy while $\mathbb{E}$ for expectation.

$$I(C, z) = \mathbb{H}(C) - \mathbb{H}(C|z)$$

$$= \mathbb{H}(C) + \mathbb{E}_{z \sim p(z|C)}[\mathbb{E}_{\hat{C} \sim p(\hat{C}|z)}[\log p(\hat{C})]]$$

$$= \mathbb{H}(C) + \mathbb{E}_{C \sim p(C)}[DKL[p(\hat{C})||p_{\phi}(\hat{C})] + \mathbb{E}_{\hat{C} \sim p(\hat{C}|z)}[\log p(\hat{C})]]$$

$$\geq \mathbb{H}(C) + \mathbb{E}_{C \sim p(C)}[\mathbb{E}_{\hat{C} \sim p(\hat{C}|z)}[\log p_{\phi}(\hat{C})]]$$

We similarly compute the expression for $I(i, z)$:

$$I(i, z) \geq \mathbb{H}(i) + \mathbb{E}_{z \sim p(z|i)}[\mathbb{E}_{\hat{z} \sim p(\hat{z}|i)}[\log p_{\phi}(\hat{z})]]$$

The expression for $I(q, z|i, C)$ then follows as:

$$I(q, z|i, C) \geq \mathbb{H}(q) + \mathbb{E}_{q \sim p(q|i)}[\mathbb{E}_{z \sim p(z|i, C)}[\log p_{\phi}(\hat{z}|i, C)]]$$

where $p(q|z, i, C) = p(q|z)p(z|i, C)$

We substitute equations 4, 5, and 6 in equation 1:

$$\max_{\phi} \mathbb{E}_{p_{\phi}(q, i, C)}[\log p_{\phi}(q|z, i, C) + \lambda_1 \log p_{\phi}(i|z) + \lambda_2 \log p_{\phi}(C|i, z)]$$

where $p_{\phi}(q, i, C) = p_{\phi}(q|z)p_{\phi}(z|i, C)p_{\phi}(i|C)$

Hence, we can optimize the variational lower bound by maximizing the image and category reconstruction whilst also maximizing the MLE of question generation.

2 INFERENCE FRAMEWORK
We illustrate the inference flow using Figure 1. During inference, given an image conditioned on the category label, $z_i$ is sampled from the combined generative latent representation $z$ of the inputs learnt by the model. This representation is then passed through the temporal network $LSTM_q$, thereby, outputting the generated question. While, the training of C3VQG requires images and their corresponding ground-truth questions from different answer categories, the inference only requires the images with answer categories of the questions to be generated.

3 HYPERPARAMETERS
We present all a list of all the hyperparameter values used in training the C3VQG model.

| Hyperparameter                  | Symbol | Value |
|--------------------------------|--------|-------|
| Image Recon. Weight            | $\lambda_I$ | 1.0   |
| Category Recon. Weight         | $\lambda_C$ | 2.0   |
| Question Recon. Weight         | $\lambda_Q$ | 3.0   |
| Category Consistency Weight    | $\lambda_{cons}$ | 2.0   |
| Center Loss Weight             | $\lambda_{center}$ | 3.0   |
| Hyper-prior KL-Divergence Weight | $\lambda_{bhv}$ | 3.0   |
| Hyper-prior Regularisation Weight | $\lambda_{reg}$ | 2.0   |
| Dimension of combined latent space | $d$ | 64    |
| Learning Rate                  | $\alpha_{LR}$ | 1e-3  |

Table 1: Hyperparameters values used for training C3VQG.

4 DATASET DETAILS
We list the details about the VQA dataset [?] used for the training and evaluation of C3VQG against the state-of-the-art [? ? ?] in VOG.

| Data Type                      | Training | Validation |
|-------------------------------|----------|------------|
| VQA Annotations (answers)     | 4,437,570| 2,143,540  |
| VQA Input Questions           | 443,757  | 214,354    |
| VQA Input Images              | 82,783   | 40,504     |

Table 2: Dataset details for the VQA dataset.
Figure 1: C3VQG Inference Framework