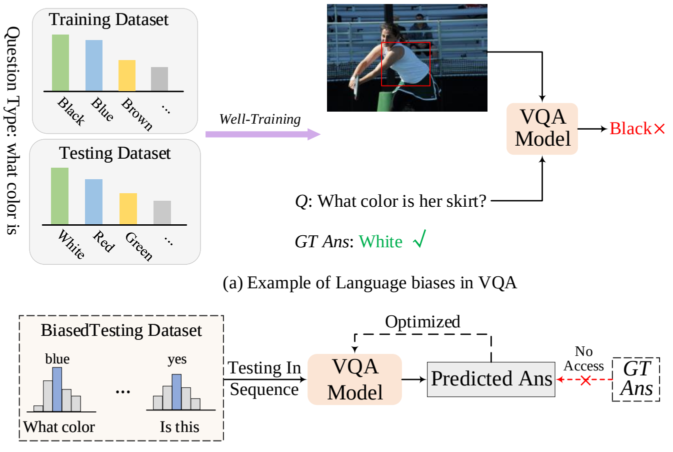
Question Type-Aware Debiasing for Test-time Visual Question Answering Model Adaptation

Jin Liu, Jialong Xie, Fengyu Zhou, and Shengfeng He, *Senior Member, IEEE*

***Abstract*—In Visual Question Answering (VQA), addressing language prior bias, where models excessively rely on superficial correlations between questions and answers, is crucial. This issue becomes more pronounced in real-world applications with diverse domains and varied question-answer distributions during testing. To tackle this challenge, Test-time Adaptation (TTA) has emerged, allowing pre-trained VQA models to adapt using unlabeled test samples. Current state-of-the-art models select reliable test samples based on fixed entropy thresholds and employ self-supervised debiasing techniques. However, these methods struggle with diverse answer spaces linked to different question types and may fail to identify biased samples that still leverage relevant visual context. In this paper, we propose Question type-guided Entropy Minimization and Debiasing (QED) as a solution for test-time VQA model adaptation. Our approach involves adaptive entropy minimization based on question types to improve the identification of fine-grained and unreliable samples. Additionally, we generate negative samples for each test sample and label them as biased if their answer entropy change rate significantly differs from positive test samples, subsequently removing them. We evaluate our approach on two public benchmarks, VQA-CP v2, and VQA-CP v1, and achieve new state-of-the-art results, with overall accuracy rates of 48.13% and 46.18%, respectively.**

***Index Terms*—Test-time adaptation, visual question answering, language debiasing**



(a) Example of Language biases in VQA

(b) Example of Test time adaption in VQA

Fig. 1. Test-time Adaptation for Visual Question Answering with Biased Dataset Distributions. (a) Illustration of biased training and testing subsets related to the question type "what color is" in the VQA-CP v2 [1] dataset. In the training subset, "black" constitutes a significant portion of the answers, whereas in the testing subset, "black" represents a small proportion. Current methods tend to predict the biased answer by capturing language biases within the training dataset. (b) Given the biased testing dataset and a pre-trained biased VQA model (*e.g.*, UpDn [2]), test-time adaptation aims to enhance the VQA model's out-of-distribution performance by leveraging unlabeled sequential testing samples.

I. Introduction

Visual question answering (VQA) is a prevalent and challenging multi-modal task that demands a strong grasp of visual context understanding [29] and linguistically-aware reasoning [6], [23], [45]. With the advancement of deep neural networks, VQA models have made significant strides in applications like human-robot interaction [44] and visual dialogs [46]. Typically, existing VQA models excel when they have access to large-scale training datasets and share similar data distributions between training and testing sets. However, these well-trained VQA models often struggle when faced with out-of-distribution scenarios in real-world applications [42]. In these cases, answer distributions diverge between the training and testing datasets. For instance, as depicted in Figure 1 (a), when we query a well-trained VQA model with "What color is her shirt?", it might provide a biased answer like "black", reflecting the answer distribution in the training dataset, while neglecting the true visual context. Therefore, it is crucial to appropriately adapt the deployed VQA model in one distinct scene to ensure optimal performance when facing distribution shifts in test samples.

Recently, researchers have explored the utility of test-time adaptation, a well-established technique in the field of image classification [26], [35], to address the model adaptation issues [38]. Figure 1 (b) outlines a simplified workflow for test-time adaptation. This process optimizes model parameters exclusively using unlabeled test samples. In previous test-time adaptation approaches, entropy minimization has been a common strategy to identify unreliable samples and mitigate

This work was supported in part by the Key R & D Project of Shandong Province under Grant 2023TZXD018, the Jinan "20 New Colleges and Universities" Funded Scientific Research Leader Studio under Grant 2021GXRC079, the Central Government Guiding Local Science and Technology Development Foundation of Shandong Province under Grant YDZX2023122, the Jinan City and University Cooperation Development Strategy Project under Grant JNSX2023012, the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant 2023B1515020097, and National Research Foundation Singapore under the AI Singapore Programme under Grant AISG3-GV-2023-011. Besides, this research was supported by Meituan Academy of Robotics Shenzhen.

Jin Liu, Jialong Xie, and Fengyu Zhou are with the School of Control Science and Engineering, Shandong University, China. Email: {202120638, 202220703}@mail.sdu.edu.cn, zhoufengyu@sdu.edu.cn. Shengfeng He is with the School of Computing and Information Systems, Singapore Management University, Singapore. Email: shengfenghe@smu.edu.sg.

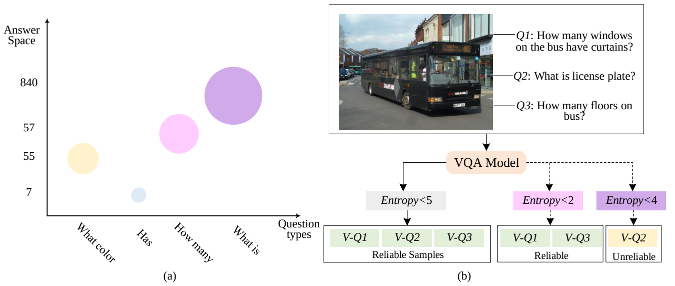


Fig. 2. (a) Questions of diverse question types are associated with varying answer spaces and entropy values. Larger circle sizes indicate higher entropy values. (b) Utilizing a dynamic entropy filter based on question types enhances the precision of sample selection, leading to a more fine-grained and reliable process.

the gradient effects stemming from them [26]. However, as highlighted by Wen et al. [38], directly applying these methods to VQA tasks often results in unsatisfactory performance for two main reasons. Firstly, these methods tend to overlook language distribution bias, as illustrated in Figure 1(a), leading to suboptimal optimization during test-time adaptation. Secondly, pre-trained models like UpDn [2] inherently capture language biases and cannot be directly optimized without considering bias prediction effects.

To tackle the above challenges, Wen et al. [38] proposed a strategy to mitigate bias during test-time adaptation by minimizing entropy for biased and reliable test samples in VQA model adaptation. However, their method relies on a fixed entropy threshold to identify unreliable samples, as illustrated in Figure 2(a), leading to an inflexible sample selection process. This raises critical questions about effectively addressing bias issues during test-time adaptation, especially when questions of different types lead to varying answer spaces and entropy values. For instance, "how many" questions typically yield numerical answers from a limited set of candidates, resulting in lower entropy values compared to "what is" questions. Additionally, we face the problem of existing VQA models, such as UpDn [2], often predicting biased answers while they still utilize the visual context. For instance, as shown in the example on the left side of Figure 3, when asked whether there is a man or woman in the picture, the VQA model predicts "Yes" due to data distribution bias. As the example on the right side, the model provides "No" for the question regarding the actual visual context. The conventional approach of distinguishing biased samples by comparing top-1 predicted answers between positive and negative samples proves unreliable, as their inconsistent predictions may be attributed to changing visual content.

To comprehensively address these challenges, we present Question Type-guided Entropy Minimization and Debiasing (QED), a two-fold approach aimed at enhancing test-time adaptation in Visual Question Answering (VQA). Our first solution acknowledges the significant influence of question types on answer spaces and entropy values. To adapt dynamically to different question types, we introduce a dynamic entropy filter derived from the correlation between answers and question types. This innovative mechanism, depicted in Figure 2(b), facilitates the filtration of entropy values during test-time adaptation. By effectively identifying unreliable samples, it optimizes the model's performance in evolving data distributions. Leveraging the identified reliable samples, we can further refine the language debiasing process.

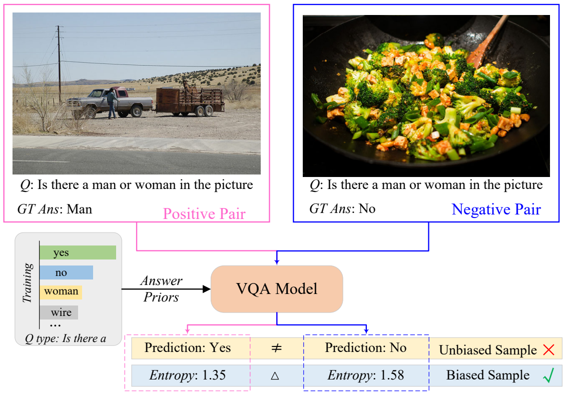


Fig. 3. Example of answer entropy change rate for identifying the biased and unbiased test samples. The existing approach [38] considers the aforementioned sample to be unbiased (the answer is wrong ✗), given that the VQA model generates distinct output responses for both positive and negative samples. Nevertheless, our model measures whether it is a biased sample by the change in entropy value, and regards the above sample is biased as entropy changing rate is smaller than the threshold δ (the answer is right ✓).

Our second solution addresses the challenge of identifying biased samples. Instead of solely relying on comparing top-1 predicted answers between positive and negative samples, we introduce the concept of the answer entropy change rate. This is achieved by generating negative samples for each test sample and calculating entropy values for both negative and positive samples. The answer entropy change rate quantifies the rate of change in answer entropy, enabling us to accurately identify biased test samples when this rate is insignificant. As exemplified in Figure 3, this refinement proves effective in identifying biased samples, ultimately leading to significant performance enhancements on the VQA-CP v1 and VQA-CP v2 datasets.

In summary, our contributions are as follows:

- We introduce an innovative approach that incorporates adaptive entropy minimization based on question types, enabling fine-grained and unreliable sample identification during test-time adaptation.
- We propose a novel metric, the answer entropy change rate, which effectively identifies biased test samples and enhances model performance, particularly in out-of-distribution (OOD) scenarios.
- Extensive experiments conducted on two benchmark datasets, VQA-CP v1 [23] and VQA-CP v2 [1], demonstrate the robustness and generalization capabilities of our proposed model.

Upon acceptance, we will make the source code for our approach publicly available, contributing to the broader research community.

II. Related Work

In this section, we will discuss two highly relevant studies to our work, *i.e.*, Overcoming language bias in visual question answering and test-time adaptation.

A. Overcoming Language Bias in Visual Question Answering

Visual question answering (VQA), introduced in [4], plays an important role in assisting visually impaired people [5], [37], requiring an intelligent agent to provide coherent answers according to the given image [8]. Despite achieving impressive results, recent studies have highlighted that current VQA models heavily rely on language distribution priors for generating answers. Once the language distribution is distinct between the training and testing datasets, current VQA models suffer from significant performance degradation due to such language bias. To enhance the robustness of existing VQA models, researchers focus on three mainstream approaches, *i.e.*, data construction, data augmentation, and model design. As for the methods with data construction, researchers focus on designing diagnosing or balanced datasets to ensure a clear distinction in answer distribution between the training and testing datasets, thereby aiding the model in mitigating language distribution bias [1], [4], [15]. For example, Agrawal et al. [15] propose to re-split the VQA v2 dataset to obtain the diagnosing dataset VQA-CP v2 with imbalanced data distributions. As can be expected, existing models that do not account for bias issues drop dramatically on the VQA-CP v2 datasets. Conditioned such dataset, another branch directly augments the dataset with extra human annotations to make the training process more balanced [7], [14], [31], [41]. For example, Gokhale et al. [14] propose to generate mutant samples by manipulating the image or the questions to expand the training sample space. Nevertheless, these methods typically require extra human annotations, which may be costly and difficult for researchers to collect suitable samples. To mitigate the dependency on the extra annotations, recent methods propose to construct negative samples with self-supervised learning [9], [17], [18], [47] or counterfactual-based learning paradigms [20], [27], [40]. For example, Cho [17] et al. propose a generative method to train the bias model directly from the target model to distinguish the language bias rather than introducing extra branch. Huai [9] et al. propose to constructs negative samples by randomly replacing the question types of the samples to formulate a self-supervised method to mitigate the question-to-answer bias without using external annotations. Kolling [20] et al. introduce a answer-assignment mechanism that exploits the probability distribution of the answers based on their frequencies to generate the counterfactual samples for the debiasing process. These methods typically introduce debiased techniques in the training time and select the optimal model conditioned on the performance of the testing dataset. However, in real-world applications, the model can not access the OOD distributions of the test samples until it has been evaluated [38]. Therefore, in this paper, we follow this novel and intuitive setting as our primary starting point. Here, we optimize the VQA model with unlabelled test data that have different distributions from the training data and utilizing the test-time adaptation methods to achieve such objective.

B. Test-time Adaptation

Test-time adaptation techniques [32], [35], [36], aiming to adapt a pre-trained model to a potentially shifted target test domain without test label information, are widely utilized in the image classification field. For example, Sun et al. [33] first propose the general training pipeline of test-time adaptation. Specifically, a pre-trained classification model is first given during the test-time optimization. Then, the author trains the model with a test sample with an auxiliary self-supervised rotation prediction task. After being fully trained, the updated model will be utilized for the final prediction. After that, some researchers adopt parameter regularization approaches to avoid altering the training process and improve adaptation efficiency [13], [26]. For example, Niu et al. [26] propose minimizing the entropy loss with reliable and non-redundant samples selected by an active sample selection criterion for test-time adaptation. Subsequently, the concept of test-time adaptation has garnered significant attention and found applications in various real-world scenarios, such as object recognition [19], human pose estimation [21], and so on. Regrettably, to date, only one work has concentrated on comprehensive test-time adaptation in the context of VQA [38], which is more in line with real-world application scenarios. In [38], the authors account for the out-of-distribution nature of the debiasing visual question answering (VQA) problem and then propose to minimize the entropy for the biased and unreliable test samples for VQA model adaptation. However, they overlook the fact that questions of different types can have diverse answer spaces. Applying a fixed entropy threshold may not suit ever-changing data distributions. In contrast, we propose a question type-guided entropy minimization and language debiasing approach that fully utilizes dynamic entropy filter constraints for adapting VQA models during test time.

III. Methods

In this section, we first formulate the task definition of test-time adaptation for visual question answering and then elaborately introduce the proposed model.

A. Task Definition

Following previous work [38], in this paper, we aim to adapt a pre-trained VQA model at test time, where the model can only access the unlabelled test data. Notably, the setting is still conditioned on the general VQA task definition, which is also regarded as a multi-classification task. In specific, given a training dataset $\mathcal{D}_{tr} = \{(v_i, q_i, a_i)\}_{i=1}^{N_{tr}}$ consisting of N_{tr} triplets of one image v_i, one question q_i and one corresponding answer $a_i \in \mathcal{A}$, we first pre-train a VQA model f with parameter θ by minimizing the loss between the prediction and ground truth, as described by the following equation.

$$\min_{\theta} - \frac{1}{N_{tr}} \sum_{i}^{N_{tr}} \mathcal{L}(f(v_i, q_i; \theta), a_i), \tag{1}$$

where $\mathcal{L}(\cdot)$ denotes the loss function. After the model is fully trained, we seek to adapt it at test time to improve the model performance given the test dataset $\mathcal{D}_{te} = \{(v_j, q_j)\}_{j=1}^{N_{te}}$ without any test label information. The process can be formulated as:

$$\min_{\theta} -\frac{1}{N_{te}}\sum_{i}^{N_{te}} \mathcal{L}(f(v_j, q_j; \theta)), \tag{2}$$

where N_{te} denotes the number of triplets in the testing dataset.

B. Our Proposed Method

In the task of test-time adaptation for visual question answering, two important issues should be fully considered. 1) as pointed out in [26], [38], the samples with high entropy may result in the wrong gradient update direction for the VQA model. Current methods utilize fixed thresholds to filter the unreliable samples with high predicted entropy. We argue that questions with different question types can own diverse answer spaces, and simply applying a fixed entropy threshold to construct filtering masks may not be suitable for ever-changing data distributions. Therefore, we introduce adaptive entropy minimization according to the question types for fine-grained and unreliable sample identification processes. 2) The language bias is ubiquitous in the VQA datasets [10], [22], existing VQA models inevitably capture the language biases instead of truly reasoning over the questions. To identify the biased sample, current methods typically construct negative and positive question-image pairs and subsequently compare the top-1 answers between these two predictions. However, recent studies reveal that though some VQA models have biased predictions [23], [39], they still utilize certain visual contexts. Therefore, some biased samples can not be distinguished by comparing two predicted top-1 candidate answers from positive and negative samples, respectively. In contrast, we generate negative samples for each test sample and identify the test samples as biased if their answer entropy change rate is not significant when compared to each positive test sample. In overall, our proposed QED model is mainly designed to remove unreliable and biased samples for test-time adaptation and improve the out-of-distribution performance. The overall algorithm is shown in Alg. 1. Sequentially, we introduce each component in our proposed model in detail.

Algorithm 1 The pipeline of the proposed model.

```
Require: The pre-trained VQA model $f$ with parameter $\theta$.
    Test samples $\mathcal{D}_{te} = \{(v_j, q_j)\}_{j=1}^{N_{te}}$. The batch size $B$. The
    obtained answer space $AS^{qt}$.
Ensure: The predictions $\{\hat{a}\}_{j=1}^{N_{te}}$ and the model $f$
for sample a batch data $D_j = \{(v_j, q_j)\}_{j=1}^{B}$ from $\mathcal{D}_{te}$ do
    Obtain the predictions $\hat{a}_j$ and entropy via Eq.(5)
    Identify the unreliable samples via Eq.(6)
    Construct negative samples $\tilde{D}_j = \{(\tilde{v}_j, q_j)\}_{j=1}^{B}$
    Calculate entropy for positive and negative samples via
        Eq.(7) and Eq.(8).
    Identify the biased samples via Eq.(9)
    Update the model $f$ with Eq.(13)
end for
```

1) Dynamic Entropy Filtering: As discussed above, samples with high entropy reveal uncertain predictions and may lead to noise gradients [26], [38]. Considering the fact that questions with different question types own distinct answer spaces and varying inherent entropy values, in this paper, we introduce adaptive entropy minimization according to the question types for a fine-grained and unreliable sample identification process.

In specific, we first collect the question types (QT) from VQA v2 [15] for efficient calculation during the test-time adaptation. Notably, we didn't utilize any other data for pre-training the VQA model. Subsequently, we directly collect the answers without any extra question information and obtain the answer space AS over the whole dataset and obtain a original answer space AS^{qt} corresponding to the question types. Then, we utilize a pre-trained BERT [12] to judge the correlations between specific question type qt and different answers by the following equation:

$$corr_{qt} = \underset{qt \in QT}{\text{softmax}}(\text{BERT}(qt) * \text{BERT}(AS)) \in \mathcal{R}^{|AS|}, \tag{3}$$

where the symbol * denotes the operation for calculating similarity. BERT$(\cdot)$ denotes the language modeling function by BERT. Next, we define a refined answer space tailored to specific question types by selecting the top answers based on their correlation values. In specific, Conditioned on the scores, we retained the top-1000 most relevant results. Next, we intersect the answer set AS^{qt} with the sorted results to obtain a new answer space AS^{nqt} for different question type qt. In contrast to the approach proposed by Wen et al. [38], who set a fixed entropy threshold to select reliable samples, we introduce a dynamic entropy filter according to the question types for fine-grained and unreliable sample filtering. The dynamic entropy filtering process is illustrated in Figure 4 and can be formulated as follows:

$$T_{qt}^{rel}(v_j, q_j) = \mathbb{I}_{\{Entropy(f(v_j,q_j;\theta))<Ent(|AS^{qt}|)\}}(v_j, q_j) \tag{4}$$

where T_{qt}^{rel} denotes the reliable sample with question type qt. $\mathbb{I}$ denotes the indicator function, where the value is 1 (*i.e.*, reliable sample) when the entropy of sample (v_j, q_j) is smaller than the dynamic entropy filter $Ent(|AS^{qt}|)$, and 0 otherwise. Notably, following previous work [38], the dynamic entropy filter is further adjusted via a hyper-paramter α, *i.e.*, $Ent(|AS^{qt}|) = \alpha \ln|AS^{qt}|,\ \alpha \in [0, 1]$. $|AS^{qt}|$ denotes the number of candidate answers for the question type qt. $Entropy(\cdot)$ is the typical entropy calculation equation for the sample (v_j, q_j) conditioned on the predictions of the VQA model [26], [38], which can be obtained by the following equation:

$$Entropy(\hat{a}_j) = -\frac{1}{N_{te}}\sum_{j=1}^{N_{te}} \text{softmax}(\hat{a}_j) * \text{log\_softmax}(\hat{a}_j),$$
$$\hat{a}_j = f(v_j, q_j; \theta)\ j \in 1, \cdots, N_{te}, \tag{5}$$

where log_softmax$(\cdot)$ denotes the logit calculation function. $\hat{a}_j$ represents the prediction.Conditioned on the designed adaptive

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2024.3410041

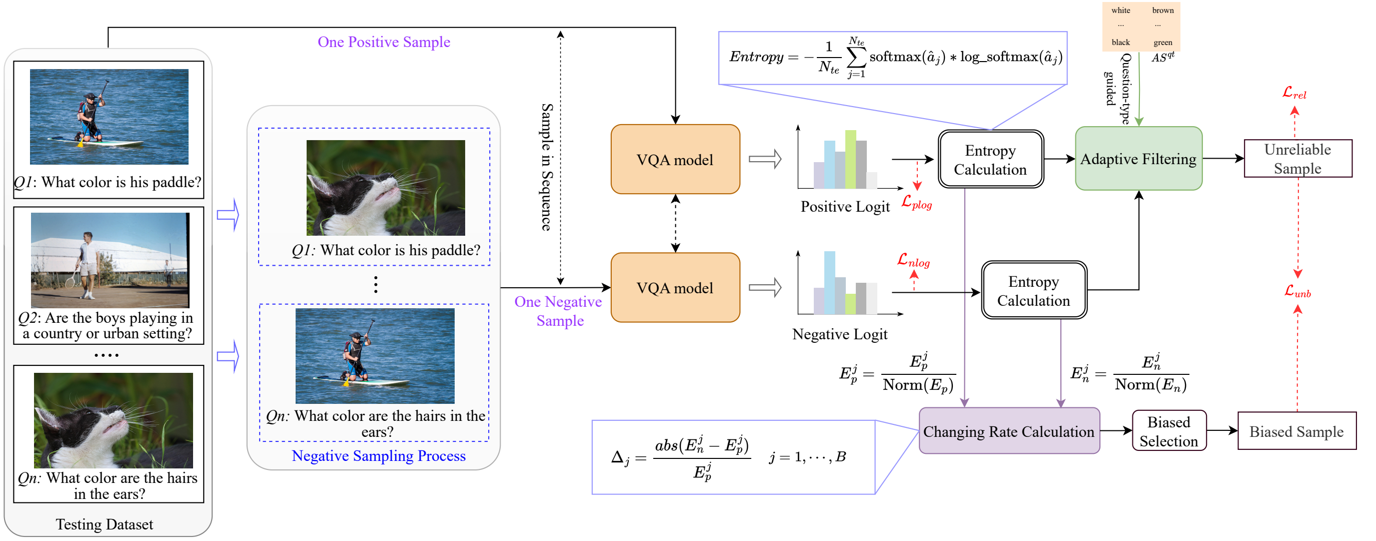

One Positive Sample
Q1: What color is his paddle?
Q2: Are the boys playing in a country or urban setting?
Qn: What color are the hairs in the ears?
Testing Dataset
Q1: What color is his paddle?
Qn: What color are the hairs in the ears?
Negative Sampling Process
Sample in Sequence
One Negative Sample
VQA model
VQA model
Positive Logit
Negative Logit
$\mathcal{L}_{plog}$
$\mathcal{L}_{nlog}$
$Entropy = -\frac{1}{N_{te}}\sum_{j=1}^{N_{te}} \text{softmax}(\hat{a}_j) * \text{log\_softmax}(\hat{a}_j)$
Entropy Calculation
Entropy Calculation
white
brown
black
green
AS^{qt}
Question-type guided
Adaptive Filtering
Unreliable Sample
$\mathcal{L}_{rel}$
$\mathcal{L}_{unb}$
$E_p^j = \frac{E_p^j}{\text{Norm}(E_p)}$
$E_n^j = \frac{E_n^j}{\text{Norm}(E_n)}$
$\Delta_j = \frac{abs(E_n^j - E_p^j)}{E_p^j} \quad j = 1, \cdots, B$
Changing Rate Calculation
Biased Selection
Biased Sample

Fig. 4. The overview of our proposed QED model. Given a pre-trained model f with parameters θ, we aim to adapt f at test time, where the model can only access the unlabelled test data. In overall, our proposed QED model is mainly designed to identify unreliable samples by a dynamic entropy filter and biased samples by entropy changing rate for test-time adaptation and the out-of-distribution debiasing.

filtering, we can identify the unreliable samples and utilize the reliable ones to update the parameters of the VQA model for test-time adaptation. The optimization loss is formulated as:

$$\mathcal{L}_{rel} = \frac{1}{N_{te}} \sum_{j=1}^{N_{te}} Entropy(f(v_j, q_j; \theta)) \cdot T_{qt}^{rel}(v_j, q_j) \qquad (6)$$

2) Entropy Change Rate: As pointed out in [16], [23], inherent language biases within the dataset lead the VQA model to rely on superficial correlations between questions and answers rather than incorporating visual information for reasoning. This issue may be exacerbated in scenarios involving test-time adaptation. Typically, an intrinsic way to alleviate the bias without utilizing extra information involves the introduction of self-supervised learning objectives. In this paper, we follow such a general pipeline.

Specifically, given a pair of test samples (v_j, q_j) that can be regarded as one positive sample, we then randomly select an irrelevant image from the mini-batch for question q_j to construct a negative sample $(\tilde{v}_j, q_j)$. Intuitively, when feeding one positive sample and its counterpart negative sample to the VQA model, two distinct answer predictions, $\hat{a}_j^{pos}$ and $\hat{a}_j^{neg}$, should be obtained. If the top-1 answer from the above two answer predictions is the same, the sample is regarded as biased, otherwise as unbiased. However, some biased samples can not be distinguished by simply comparing two predicted top-1 predictions due to the visual context utilization as discussed in Figure 3. In contrast, we propose an entropy change rate to select biased samples. Specifically, conditioned on the predictions $\hat{a}_j^{pos}$ and $\hat{a}_j^{neg}$ obtained through contrastive negative sampling, we first calculate their entropy by Equation 5 and get the value for positive prediction E_j^p and negative prediction E_j^n,

$$E_j^p = Entropy(\hat{a}_j^{pos}), \quad E_j^n = Entropy(\hat{a}_j^{neg}) \qquad (7)$$

Then, we normalize the entropy value within the mini-batch size B to ensure that the data is not overly discrete:

$$E_p^j = \frac{E_p^j}{\text{Norm}(E_p)}, \quad E_n^j = \frac{E_n^j}{\text{Norm}(E_n)}, \quad j = 1, \cdots, B, \qquad (8)$$

where Norm($\cdot$) denotes the normalization function. To more accurately determine whether it is a biased sample, we introduce the concept of answer entropy change rate Δ. If the changing rate is below a predefined threshold δ, the sample will be regarded as biased, otherwise as unbiased. The process can be formulated as follows:

$$\begin{aligned} \Delta_j &= \frac{\text{abs}(E_n^j - E_p^j)}{E_p^j}, \quad 1, \cdots, B \\ T_{qt}^{bias}(v_j, q_j) &= \mathbb{I}_{\{\Delta_j > \delta\}}(v_j, q_j), \end{aligned} \qquad (9)$$

where abs($\cdot$) denotes the absolute function. Conditioned on the selected unbiased sample and reliable sample, we can eliminate the negative effect of these test samples during the test-time adaptation. Consequently, the debiasing entropy minimization based on Equation 6 and Equation 9 can be formulated as follows:

$$\mathcal{L}_{qt}^{bias} = \frac{1}{N_{te}} \sum_{j=1}^{N_{te}} Entropy(f(v_j, q_j; \theta)) T_{qt}^{bias}(v_j, q_j) T_{qt}^{rel}(v_j, q_j) \qquad (10)$$

C. Contrastive Bias Removing

Apart from entropy minimization for reliable biased samples by Equation 10, following previous work [38], we additionally eliminate the language bias directly for the VQA model during test-time, *i.e.*, the samples are unbiased and reliable. The process is formulated as:

$$\tilde{T}_{qt}^{bias}(v_j, q_j) = T_{qt}^{rel}(v_j, q_j)(\sim T_{qt}^{bias}(v_j, q_j)) \qquad (11)$$

Notably, we can not access the ground-truth answers of the test samples during test-time adaptation. Consequently, the position of the prediction with the highest score is selected as the pseudo label, *i.e.*, $k = \text{argmax}(\hat{a}_j^{pos})$. Therefore, the direct optimization loss for positive samples $L_{qt}^{p_{debias}}$ and negative samples $L_{qt}^{n_{debias}}$ are formulated as:

$$\begin{aligned}\mathcal{L}_{qt}^{p\_op} &= \frac{1}{N_{te}} \sum_{j=1}^{N_{te}} \tilde{T}_{qt}^{bias}(v_j, q_j)\text{softmax}(\hat{a}_j^{pos})[k] \\ \mathcal{L}_{qt}^{n\_op} &= \frac{1}{N_{te}} \sum_{j=1}^{N_{te}} \tilde{T}_{qt}^{bias}(v_j, q_j)\text{softmax}(\hat{a}_j^{neg})[k]\end{aligned} \tag{12}$$

D. Test-time Adaptation for Visual Question Answering

In a brief summary, we aim to optimize the VQA model during test time. The overall loss mainly consists of four parts, *i.e.*, $\mathcal{L}_{rel}$ in Equation 6, $\mathcal{L}_{qt}^{bias}$ in Equation 10, $L_{qt}^{p\_op}$ and $L_{qt}^{n\_op}$ in Equation 12.

$$\mathcal{L} = \mathcal{L}_{qt}^{bias} + \mathcal{L}_{qt}^{p\_op} + \mathcal{L}_{qt}^{n\_op} + \mathcal{L}_{rel} \tag{13}$$

When all the test samples are finished reading, the model's parameters are also updated, and all prediction results are output.

IV. Experiments

In this section, we first review the two public out-of-distribution benchmarks VQA-CP v1 [23] and VQA-CP v2 [1], along with the standard evaluation metric. Then, we conducted extensive experiments from quantitative and qualitative perspectives to evaluate the proposed QED performance on two benchmarks.

A. Datasets

VQA-CP V1 is proposed to diagnose the out-of-distribution (OOD) performance of the VQA model and constructed from VQA v1 [4] by re-organizing the dataset to make training and testing sets own the different answer distributions. In Specific, the VQA-CP V1 dataset consists of $\sim$ 118K images and $\sim$ 244K questions for the training subset, $\sim$ 87K images and $\sim$ 125K questions for the testing subset. Following previous work [23], [38], we utilize the official splits for VQA-CP V1.

As for the VQA-CP V2, it is similar to the VQA-CP V1 dataset and possess a large amount of samples. The dataset is constructed from VQA V2 [15] by re-splitting the training and validation datasets to obtain different answer distributions between training and testing dataset. The VQA-CP V2 possesses a bigger answer space compared with VQA-CP V1 and is widely utilized to verify the generalization ability of the VQA model. Specifically, the VQA-CP v2 dataset consists of $\sim$121K images and $\sim$438K questions for the training subset, $\sim$98K images and $\sim$220K questions for the testing subset.

More data statistics on VQA-CP V1 and VQA-CP V2 are presented in Table I. Notably, under the test-time adaptation setting, when given the pre-trained VQA model, we can only access the testing dataset without any label information (the highlighted areas in light blue in Table I) for the test-time adaptation.

TABLE I
The statistics of VQA-CP V1 [23] and VQA-CP V2 [1].

Datasets	Train(Images/Questions)	Test	Answer Space
VQA-CP V1 [23]	118K/244K	87K/125K	1691
VQA-CP V2 [1]	121K/438K	98K/220K	2274

B. Evaluation Metric

Following previous work [23], we evaluate the performance of our proposed method and the compared methods using standard VQA accuracy metric, which can be formulated as follows.

$$acc(\hat{a}) = \min(1, \frac{count(\hat{a})}{3}), \tag{14}$$

where $\hat{a}$ denotes the predicted answer. The score of ground-truth answer for specific questions is derived from ten annotators. The accuracy can be 100% once at least three workers provide the same answer as the predicted answer $\hat{a}$. Notably, the accuracy is measured on four categories, *i.e.*, *All*, *Yes/No*, *Number*, and *Other*.

C. Implementation Details

Following previous works [23], [38], [40] in the domain of debiased VQA task, for visual features, we utilize the Faster R-CNN [30], pre-trained by [3], to extract 2048-d visual features for top 36 objects in one image. As for the textual features, we initialize each word with 300-d Glove embeddings [28]. The question is truncated or padded with a maximum of 14 words and then processed by a GRU [11] of 512-d hidden vector. For a fair comparison, we directly utilize the visual features provided by [40].

Since our proposed QED is model-agnostic, we incorporate the QED into three mainstream baseline models to improve their generalization ability on VQA-CP V1 and VQA-CP V2, *i.e.*, UpDn [2], LXMERT [34], and ViLBERT [24]. For UpDn, during test-time adaptation, we utilize SGD optimizer to update the proposed model, where the momentum for SGD is set to 0.9. The learning rate is first initialized as 0.01 with a weight decay is 0.0001. The batch size is set to 512. The hyper-parameter for adjusting entropy value α is set to 0.1. The changing rate threshold δ is set to 2.0. As for LXMERT, the learning rate is first initialized as 1e-4. The batch size is set to 32. The hyper-parameter for adjusting entropy value α is set to 0.2. The changing rate threshold δ is set to 1.0. As for ViLBERT, the learning rate is first initialized as 2e-4 for VQA-CP V1 and 5e-4 for VQA-CP V2, respectively. The batch size is set to 32. The hyper-parameter for adjusting entropy value α is set to 0.2 for VQA-CP V1 and 0.05 for VQA-CP V2. The changing rate threshold δ is set to 1.5.

Notably, for a fair experimental comparison, all the pre-trained VQA models before test-time adaptation are adopted from [38]. Besides, all the experiments are conducted on a single RTX 3090 GPU and implemented with Pytorch 1.7.

D. Main Results on VQA-CP V1 and VQA-CP V2

We compare the performance of our proposed QED built upon UpDn [2] with recent state-of-the-art methods on VQA-

TABLE II
THE EXPERIMENTAL RESULTS ON VQA-CP V2 [1] AND VQA-CP V1 [4] COMPARED WITH STATE-OF-THE-ART MODELS IN TERMS OF ACCURACY (%). I-III DENOTES THE BACKBONE MODELS, TYPICAL TEST-TIME ADAPTATION MODELS, AND TEST-TIME ADAPTATION WITH LANGUAGE DEBIASING MODELS. ALL THE MODELS ARE CONDITIONED ON UPDN [2]. THE OVERALL PERFORMANCE IS HIGHLIGHTED IN LIGHT BLUE.

Categories	Model	**VQA-CP v2 test (%)**				**VQA-CP v1 test (%)**			
		All	Yes/No	Num	Other	All	Yes/No	Num	Other
I	UpDn [2]	40.94	43.87	12.90	47.09	38.78	42.78	13.38	45.01
II	TENT [35]	41.17	43.88	13.95	47.22	38.79	42.77	13.45	45.02
	TPT [32]	41.15	43.85	14.16	47.13	38.64	42.39	13.54	45.00
	CoTTA [36]	41.01	43.86	12.97	47.21	38.80	42.74	13.39	45.09
	ETA [26]	41.28	43.81	13.76	47.51	38.95	42.89	13.73	45.15
	RoTTA	38.15	42.90	13.38	43.37	40.09	43.99	12.85	45.52
	TIPI	38.12	41.05	13.48	45.11	41.42	43.58	13.90	47.30
III	TDS† [38]	46.12	62.11	10.50	47.51	45.44	58.09	13.70	45.60
	TDS [38]	46.33	62.55	10.53	47.66	45.55	58.22	13.82	45.70
	Ours	48.16	67.13	13.48	47.73	46.78	62.34	13.49	44.68

CP V1 [23] and VQA-CP V2 [1] in terms of Accuracy (Equation 14). The results are reported in Table II. Before diving into the results, we first introduce the test-time adaptation methods briefly.

- TENT [35] proposes to optimize the model by entropy minimization for the test sample at test time.
- TPT [32] utilizes adaptive prompts to learn on the fly with a single test sample for the image classification task.
- CoTTA [36] proposes to eliminate the accumulation error and catastrophic forgetting problems in the continual test-time adaptation settings.
- ETA [26] introduces a Fisher regularizer to constrain important model parameters and propose active sample selection criterion to minimize the entropy loss for test-time adaptation.
- RoTTA [43] introduces a robust batch normalization strategy and a memory bank mechanism to enhance the model's adaptability.
- TIPI [25] integrates an invariance regularizer as a surrogate loss to effectively tackle the challenges posed by varying batch sizes during test-time model adaptation.
- TDS [38] is the first to introduce task of test-time adaptation and language debiasing for VQA. And they propose to select reliable samples by filtering the prediction entropy smaller than the fixed threshold and identify biased samples by self-supervised learning.

As can be seen from Table II, we can observe that:

- Our proposed method outperforms all the compared baseline models in terms of the overall accuracy metric "All" and "Yes/No" metrics on two benchmarks VQA-CP V1 [23] and VQA-CP V2 [1]. In specific, our model surpasses the state-of-the-art TDS [38] by 1.8% on VQA-CP V2 and and 1.23% on VQA-CP V1. As for the answer types "Num" and "Other", our model can still obtain comparable performance when compared with TDS [38]. These results demonstrate the effectiveness of our proposed model for test-time adaptation in VQA tasks.
- As for results of test-time adaptation models in category II (*i.e.*, TENT [35], TPT [32], CoTTA [36], and ETA [26]), they obtain slight improvement in terms of four answer types under the Accuracy(%) metric on VQA-CP V1 and VQA-CP V2, with only one exception case of TPT [32] on VQA-CP v1 dataset. While RoTTA [43] and TIPI [25] even brought negative effects on VQA-CP v2 dataset. This may be due to the fact that they typically ignore the language bias in the VQA task and are inferior to TDS [38] and our model in category III. TDS [38] accounts for removing the samples with high entropy and eliminating the language bias at the same time. Notably, by introducing adaptive entropy minimization and entropy changing rate, our model can further improve the performance and outperform the TDS by a large margin.

TABLE III
ABLATION STUDY ON EACH COMPONENT OF OUR MODEL IN TERMS OF ACCURACY METRIC (%). WE CONDUCT EXPERIMENTS ON VQA-CP V2 [1] AND THE MODEL IS BUILT UPON UPDN [2] BACKBONE. ENTROPY CHANGING RATE (ECR) MODULE FOR ACCURATE BIASED SAMPLE SELECTION WHICH IS SUPERVISED BY $\mathcal{L}_{qt}^{bias}$, DYNAMIC ENTROPY FILTER (DEF) MODULE FOR REMOVING UNRELIABLE SAMPLES FLEXIBLY WHICH IS SUPERVISED BY $\mathcal{L}_{rel}$, CONTRASTIVE BIAS REMOVING (CBR) MODULE FOR SELECTING UNBIASED SAMPLES VIA TYPICAL SELF-SUPERVISED LEARNING, WHICH IS SUPERVISED BY $L_{qt}^{p\_op}$ AND $L_{qt}^{n\_op}$, AND NEGATIVE SAMPLING (NS) MODULE ARE GRADUALLY REMOVED TO SHOW THE CONTRIBUTIONS OF EACH MODULE OF OUR MODEL.

NS	CBR	DEF	ECR	All	Yes/No	Num	Other
✓	✓	✓	✓	48.16	67.13	13.48	47.73
✓	✓	✓		47.04	65.64	12.43	46.79
✓	✓			46.43	62.37	10.46	47.94
✓				44.50	58.26	10.18	46.71
				40.94	43.87	12.90	47.09

E. Ablation Study

In this section, we pursue to investigate the effectiveness of each designed component in our proposed model by removing each component gradually, *i.e.*, Entropy Changing Rate (ECR) module for accurate biased sample selection which is supervised by $\mathcal{L}_{qt}^{bias}$ in Equation 10, Dynamic Entropy Filter (DEF)

TABLE IV
ABLATION STUDY FOR DIFFERENT BACKBONES ON VQA-CP V2 AND VQA-CP V1 IN TERMS OF ACCURACY (%).

Model	**VQA-CP V2 test (%)**				Δ ↑	**VQA-CP V1 test (%)**				Δ ↑
	All	Yes/No	Num	Other		All	Yes/No	Num	Other	
UpDn [2]	40.94	43.87	12.90	47.09	-	38.78	42.78	13.38	45.01	-
UpDn+Ours	48.16	67.13	13.48	44.68	+7.22	46.78	62.34	13.49	44.68	+8.00
ViLBERT [24]	40.75	43.87	14.94	46.20	-	39.59	45.07	15.14	43.97	-
ViLBERT+Ours	46.22	59.57	13.29	48.26	+5.47	46.05	60.00	14.63	44.82	+6.46
LXMERT [34]	41.72	43.61	14.17	48.30	-	38.21	40.77	14.81	45.06	-
LXMERT+Ours	47.39	66.46	13.87	46.59	+5.67	43.20	53.53	14.69	44.38	+4.99

module for removing unreliable samples flexibly which is supervised by $\mathcal{L}_{rel}$ in Equation 6, Contrastive Bias Removing (CBR) module for selecting unbiased samples via typical self-supervised learning, which is supervised by $L_{qt}^{p\_op}$ and $L_{qt}^{n\_op}$ in Equation 12, and Negative Sampling (NS) module for the introduction of negative samples. Notably, when the module is removed, the corresponding supervised loss is also removed. The ablation study is conducted on VQA-CP V2 [1] and built upon the typical VQA backbone UpDn [2].

We report the experimental results in Table III and obtain the following observations:

- By directly removing the Entropy Changing Rate (ECR) module, the performance of the proposed model drop by 1.12% in “All” metric, 1.49% in “Yes/No” metric, 1.05% in “Num” metric, and 0.95% in “Other” metric. The results reveal the effectiveness of the ECR module in selecting unbiased samples for test-time adaptation.
- When further removing the Dynamic Entropy Filter (DEF) module, the results further drop in terms of “All”, “Yes/No”, and “Num”. Especially in “Yes/No” and “Num”, the performance of our model significantly drop by 3.27% and 1.97%, respectively. This is mainly due to the fact that, without reliable samples and accurate unbiased samples, the model may fail to eliminate the language bias effectively.
- As can be expected, the Contrastive Bias Removing (CBR) module and the Negative Sampling (NS) module are two important components for the test-time adaptation and debiasing process. The results drop significantly by gradually removing them. In our paper, we follow previous works [38] and keep these two designs with only slight parameter modifications.

As discussed in Section III-B1, we propose adaptive entropy minimization according to the question types for a fine-grained and unreliable sample filtering process. In this section, we conduct extra experiments to validate the effectiveness of the design on VQA-CP V1 [23] and VQA-CP V2 [1]. The results are reported in Table V. As can be observed from the results, a fixed entropy threshold can help select reliable samples for test-time adaptation and improve the model performance compared with the baseline method UpDn [2] by 5.73% on VQA-CP V2 and 5.94% on VQA-CP V1. Notably, our dynamic entropy filter helps the model to acquire one fine-grained and unreliable sample filtering process. Thus, the model performance consistently outperforms the utilization of fixed entropy threshold by a large margin (7.19% *v.s.* 5.73% on VQA-CP V2, 8.00% *v.s.* 5.94% on VQA-CP V1). These results further demonstrate the effectiveness of our adaptive entropy minimization process for test-time adaptation. To further reveal more details of the proposed QED, we conduct more ablation studies on the VQA-CP V2 dataset regarding the following questions:

TABLE V
RESULTS COMPARISON OF FIXED ENTROPY THRESHOLD AND DYNAMIC ENTROPY FILTER ON VQA-CP V2 [1] AND VQA-CP V1 [23] IN TERMS OF ACCURACY (%).

Model	**VQA-CP V2 test (%)**				Δ ↑
	All	Yes/No	Number	Other	
Baseline Model	40.94	43.87	12.90	47.09	-
+Fixed Entropy Threshold	46.67	64.24	10.67	47.35	+5.73
+Dynamic Entropy Filter	48.16	67.13	13.48	47.73	+7.19

Model	**VQA-CP V1 test (%)**				Δ ↑
	All	Yes/No	Number	Other	
Baseline Model	38.78	42.78	13.38	45.01	-
+Fixed Entropy Threshold	44.72	56.77	13.42	45.30	+5.94
+Dynamic Entropy Filter	46.78	62.34	13.49	44.68	+8.00

TABLE VI
ABLATION STUDY FOR ADJUSTED HYPER-PARAMETER α IN DYNAMIC ENTROPY FILTER ON VQA-CP V2 [1] IN TERMS OF ACCURACY (%) FOR ALL, YES/NO, NUMBER, AND OTHER ANSWER TYPES.

α	**VQA-CP V2 test (%)**				Δ ↑
	All	Yes/No	Number	Other	
0.10	48.16	67.17	13.53	47.70	-
0.15	47.94	66.67	13.43	47.59	-0.22
0.20	47.43	65.23	13.20	47.50	-0.73
0.25	46.83	64.25	12.15	47.22	-1.33
0.30	46.38	63.53	11.53	46.96	-1.78

1) Since our proposed QED is model-agnostic, can it improve the performance of other state-of-the-art VQA backbones?
2) In real-world applications, data streams typically have different sizes. How does our model perform at different batch sizes?
3) Various parameter settings may affect the model performance in deep learning tasks. How do parameter settings affect the model performance?

Q1: *Can QED improve the performance of other state-of-the-art VQA backbones?* Before diving into the results, we first introduce the backbones briefly.

TABLE VII
RESULTS OF DIFFERENT BATCH SIZE (BS) FOR MODEL PERFORMANCE ON VQA-CP V2 [1] IN TERMS OF ACCURACY (%). THE BACKBONE IS UPDN [2]

BS	VQA-CP v2				Variations
	All	Yes/No	Number	Other	
1	48.86	71.60	16.52	45.82	+0.70
2	47.91	67.55	13.66	47.02	-0.25
4	47.85	65.45	12.64	48.29	-0.31
8	47.94	65.96	12.00	48.35	-0.22
16	48.16	67.93	12.79	47.50	+0.00
32	48.02	68.24	13.13	47.01	-0.14
64	48.22	68.18	13.34	47.34	+0.06
128	48.19	67.54	13.48	47.57	+0.03
256	47.92	65.70	13.42	48.07	-0.24
512	48.16	67.13	13.48	47.73	-
1024	47.67	67.17	13.42	46.85	-0.49

- UpDn [2] is the most prevalent VQA model and obtains the highly related object features for the question answering by an object-level attention network, which is built upon the standard top-down attention mechanism.
- ViLBERT [24] utilize a two-stream interaction module to process both visual and textual inputs. Before applying task-agnostic downstream applications, the model is first pre-trained through two proxy tasks on large datasets.
- LXMERT [34] is the newly Transformer-based cross-modality modeling framework, which is pre-trained on the large-scale paired questions and images.

From the results reported in Table IV, we can observe that our proposed model can improve the overall performance (*i.e.*, "All" metric) on VQA-CP V2 [1] and VQA-CP V1 [23] test dataset regardless of the backbones. In Specific, comparing with transformer-based models ViLBERT [24] and LXMERT [34], our proposed method improves the performance of UpDn by a large margin (7.22% *v.s.* 5.67% on VQA-CP V2 and 8.00% *v.s.* 6.46% on VQA-CP V1). We speculate that the modality interaction mechanism in transformer-based models can make the model accurately learn the visual and textual representations compared with the typical fusion strategy in UpDn. These results on both two benchmarks further reveal that our method is model-agnostic and suitable for combination with other VQA backbones to address the test-time adaptation issues.

Q2: *How does our model perform at different batch sizes?* In real-world applications, data streams are consistently fed into the model at varying sizes. Consequently, to simulate such a scenario, we conduct extra experiments on VQA-CP V2 [1] with different batch sizes. The results are reported in Table VII. From the results, we can observe that:

- When the commonly used batch size changes from 2 to 1024, our model exhibits robust performance changes, with numerical variations ranging from -0.49% (minimum) to +0.06% (maximum). The results reveal that our model is robust for test-time adaptation and can be fully utilized in real-world scenarios.
- Following previous works [35], [38], we conduct experiments under a extreme condition, *i.e.*, the batch size is set to 1. The results are reported in the first row of Table VII. We can observe that our model even achieves better results than when the batch size is 512. We speculate that this may be contributed to the negative samples which are obtained from Gaussian noise as [38]. However, the results further embody the robustness of our model.
- To ensure consistency between the testing optimization process and the training process, in this paper, we set the batch size to 512, which is the same as the pre-trained model obtained process. We find that the performance still satisfies the requirements of real-world applications, further validating the effectiveness of our method in test-time adaptation.

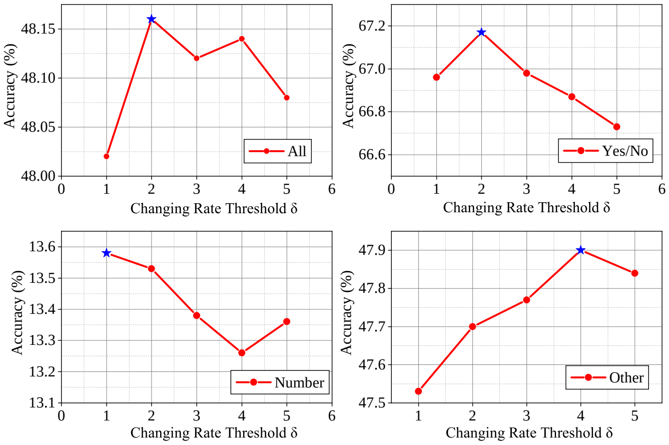


Fig. 5. Ablation study for changing rate threshold δ on VQA-CP V2 [1] in terms of Accuracy (%) for All, Yes/No, Number, and Other answer types. The blue star mark denotes the best results.

Q3: *How do parameter settings affect the model performance?* To further investigate the effectiveness of adjusted hyper-parameter α in dynamic entropy filter setting (Refer to in Section III-B1), and changing rate threshold δ for selecting biased samples in Equation 9, we conduct extensive experiments on VQA-CP V2 dataset [1]. The results are reported in Table VI and illustrated in Figure 5.

As for the adjusted hyper-parameter α in the dynamic entropy filter setting, we set α to 0.1 to 0.3 in steps of 0.05. As can be observed from the results in Table VI, when the α becomes larger, the accuracy of our proposed QED decreases gradually. This is because as α increases, many test samples with high entropy values are selected, which affects the overall performance of the model. In this paper, we set α to 0.1 on all the experiments.

To demonstrate the effectiveness of changing rate threshold δ, we set it to 1 to 5 in steps of 1. As can be seen from Figure 5, our proposed QED achieves the best performance on "All" and "Yes/No" when the δ is 2, on "Number" when the δ is 1, and on "Other" when the δ is 4. Consequently, considering the overall performance, we set the changing rate threshold δ to 2 through all the experiments for unbiased sample selection.

F. Qualitative Analysis

In this section, to further explore the effectiveness of our proposed QED on overcoming the language bias and accurate

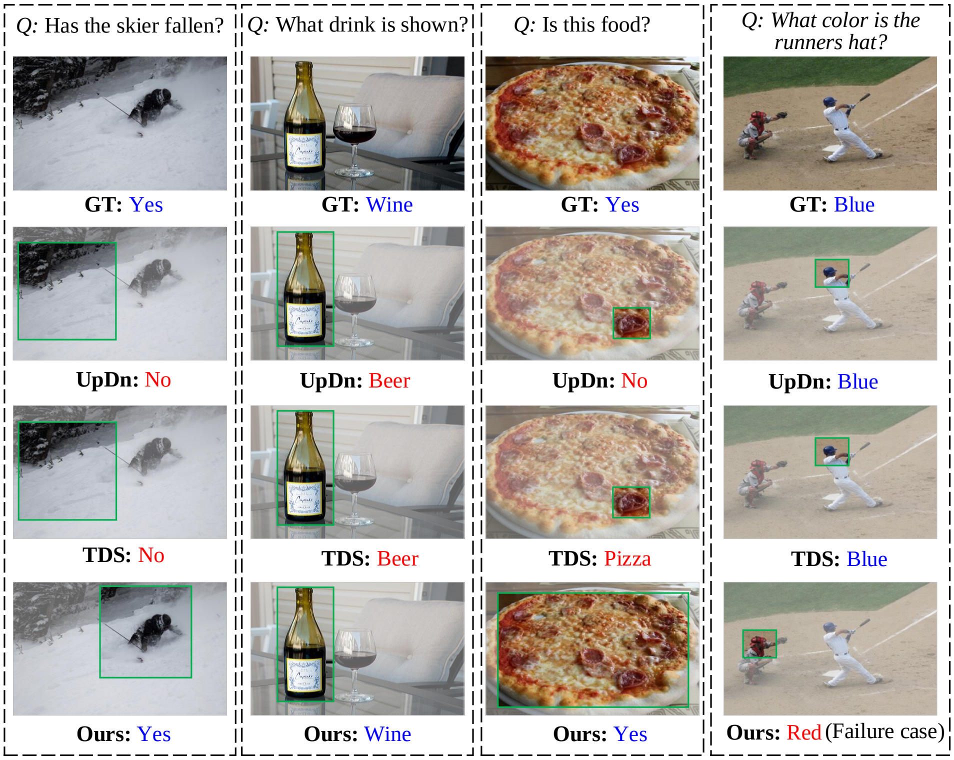


Fig. 6. Qualitative comparison results of UpDn [2] backbone, the state-of-the-art model TDS [38], and our model on VQA-CP V2 [1]. The word in blue denotes the right answer for the corresponding question. The green rectangle indicates the visual attention area from the model. GT denotes the ground-truth answer.

visual location, we illustrate four qualitative examples including one failure case in Figure 6. From the leftmost to the right, three typical biased cases are shown, *i.e.*, language bias with wrong visual location, language distribution bias, and wrong visual location.

As can be observed from the results, for the left-most column, UpDn [2] and TDS [38] fail to answer the question with incorrect visual attention location, while our proposed model can predict the right answer with appropriate visual location, revealing the effectiveness of our model in test-time adaptation and overcoming the bias. For the second case, we can find that UpDn and TDS still provide the wrong answers even they focus on the right visual attention area. This may be due to the fact that the language priors still affect the model performance and further demonstrate the effectiveness of our model. As for the third column, we can observe that our model can enhance the visual location ability of the VQA model, thereby improving the accuracy of prediction. In overall, our model can correctly predict the right answers for the questions with the right visual grounding. Notably, our model still suffers failure, which is illustrated in the rightmost column of Figure 6. We can observe that our model fails to answer the questions and provide incorrect visual grounding for “runners hat”. We speculate that this may be due to the ambiguity of the question. That’s to say, it’s hard to determine which people are the “runner”. A possible solution is to introduce a knowledge graph to assist our model in reasoning over the events for the images, and we leave it for our future work.

V. Discussion

As can be observed from the results reported in the Experiment section, our proposed model obtains significant improvements in the ”Yes/No” metric, with less impressive results in the ”Number” and ”Other” metrics. The reasons for this phenomenon can be divided into two aspects: 1) The most critical issue is due to the significant differences in data distribution among different types of questions in the test dataset. E.g., in the test dataset of VQA-CP v2, the testing data size for ”other” type is more than 100000 with more than 1000 kinds of answers, for ”number” is more than 30000 with more than 1000 kinds of answers, while for ”yes/no” is 60000 with less than 5 kinds of answers. There are significant differences in the scale of these data. Consequently, the answer for ”yes/no” is easier to obtain and more accurate compared to the other two situations. 2) Our debiasing test-time adaptation model is built upon the typical VQA backbones (e.g., UpDn and LXMERT). Nevertheless, the inference ability of these VQA backbones is relatively weak, resulting in poor performance in the ”Num” metric. One possible solution is to introduce LLMs to retrieve more relevant knowledge for our model, thereby enhancing its reasoning ability. We leave it for future work.

VI. Conclusion

In this paper, we focus on the test-time adaptation and language debiasing for VQA models, using only test samples without any label information. Conditioned on the general self-supervised sampling pipeline, we first introduce adaptive entropy minimization according to the question types to remove the unreliable samples. Then, to mitigate language bias, we propose an answer entropy changing rate. Specifically, we generate negative samples for each test sample and identify test samples as biased if their change rate is not significant when compared to positive test samples, and then remove the biased samples. Extensive experiments on VQA-CP v2 and VQA-CP v1 demonstrate that our model achieves new state-of-the-art results in overall accuracy.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980, 2018.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[5] Silvio Barra, Carmen Bisogni, Maria De Marsico, and Stefano Ricciardi. Visual question answering: Which investigated applications? *Pattern Recognition Letters*, 151:325–331, 2021.

[6] Yandong Bi, Huajie Jiang, Yongli Hu, Yanfeng Sun, and Baocai Yin. See and learn more: Dense caption-aware representation for visual question answering. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[7] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10800–10809, 2020.

[8] Zailong Chen, Lei Wang, Peng Wang, and Peng Gao. Question-aware global-local video understanding network for audio-visual question answering. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[9] Jae Won Cho, Dong-Jin Kim, Hyeonggon Ryu, and In So Kweon. Generative bias for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11681–11690, 2023.

[10] Jae Won Cho, Dong-Jin Kim, Hyeonggon Ryu, and In So Kweon. Generative bias for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11681–11690, 2023.

[11] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] Jin Gao, Jialing Zhang, Xihui Liu, Trevor Darrell, Evan Shelhamer, and Dequan Wang. Back to the source: Diffusion-driven test-time adaptation. *arXiv preprint arXiv:2207.03442*, 2022.

[14] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Mutant: A training paradigm for out-of-distribution generalization in visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 878–892, 2020.

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[16] Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian. Greedy gradient ensemble for robust visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1584–1593, 2021.

[17] Tianyu Huai, Shuwen Yang, Junhang Zhang, Jiabao Zhao, and Liang He. Debiased visual question answering via the perspective of question types. *Pattern Recognition Letters*, 178:181–187, 2024.

[18] Yash Kant, Abhinav Moudgil, Dhruv Batra, Devi Parikh, and Harsh Agrawal. Contrast and classify: Training robust vqa models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1604–1613, 2021.

[19] Junho Kim, Inwoo Hwang, and Young Min Kim. Ev-tta: Test-time adaptation for event-based object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17745–17754, 2022.

[20] Camila Kolling, Martin More, Nathan Gavenski, Eduardo Pooch, Otávio Parraga, and Rodrigo C Barros. Efficient counterfactual debiasing for visual question answering. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3001–3010, 2022.

[21] Yizhuo Li, Miao Hao, Zonglin Di, Nitesh Bharadwaj Gundavarapu, and Xiaolong Wang. Test-time personalization with a transformer for human pose estimation. *Advances in Neural Information Processing Systems*, 34:2583–2597, 2021.

[22] Jin Liu, ChongFeng Fan, Fengyu Zhou, and Huijuan Xu. Be flexible! learn to debias by sampling and prompting for robust visual question answering. *Information Processing & Management*, page 103296, 2023.

[23] Jin Liu, GuoXiang Wang, ChongFeng Fan, Fengyu Zhou, and HuiJuan Xu. Question-conditioned debiasing with focal visual context fusion for visual question answering. *Knowledge-Based Systems*, page 110879, 2023.

[24] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[25] A Tuan Nguyen, Thanh Nguyen-Tang, Ser-Nam Lim, and Philip HS Torr. Tipi: Test time adaptation with transformation invariance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24162–24171, 2023.

[26] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pages 16888–16905. PMLR, 2022.

[27] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710, 2021.

[28] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[29] Yansheng Qiu, Ziyuan Zhao, Hongdou Yao, Delin Chen, and Zheng Wang. Modal-aware visual prompting for incomplete multi-modal brain tumor segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3228–3239, 2023.

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[31] Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2591–2600, 2019.

[32] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

[33] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020.

[34] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5100–5111, 2019.

[35] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

[36] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-

time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022.

[37] Tian Wang, Jiakun Li, Zhaoning Kong, Xin Liu, Hichem Snoussi, and Hongqiang Lv. Digital twin improved via visual question answering for vision-language interactive mode in human–machine collaboration. *Journal of Manufacturing Systems*, 58:261–269, 2021.

[38] Zhiquan Wen, Shuaicheng Niu, Ge Li, Qingyao Wu, Mingkui Tan, and Qi Wu. Test-time model adaptation for visual question answering with debiased self-supervisions. *IEEE Transactions on Multimedia*, 2023.

[39] Zhiquan Wen, Yaowei Wang, Mingkui Tan, Qingyao Wu, and Qi Wu. Digging out discrimination information from generated samples for robust visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6910–6928, 2023.

[40] Zhiquan Wen, Guanghui Xu, Mingkui Tan, Qingyao Wu, and Qi Wu. Debiased visual question answering from feature and sample perspectives. *Advances in Neural Information Processing Systems*, 34:3784–3796, 2021.

[41] Jialin Wu and Raymond Mooney. Self-critical reasoning for robust visual question answering. *Advances in Neural Information Processing Systems*, 32, 2019.

[42] Zhengwei Yang, Meng Lin, Xian Zhong, Yu Wu, and Zheng Wang. Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1472–1481, 2023.

[43] Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15922–15932, 2023.

[44] Shengchen Zhang and Xiaohua Sun. Designing social interactions for learning personalized knowledge in service robots. In *International Conference on Human-Computer Interaction*, pages 656–671. Springer, 2022.

[45] Lichen Zhao, Daigang Cai, Jing Zhang, Lu Sheng, Dong Xu, Rui Zheng, Yinjie Zhao, Lipeng Wang, and Xibo Fan. Towards explainable 3d grounded visual question answering: A new benchmark and strong baseline. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

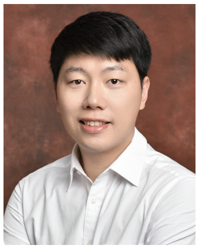
[46] Lei Zhao, Junlin Li, Lianli Gao, Yunbo Rao, Jingkuan Song, and Heng Tao Shen. Heterogeneous knowledge network for visual dialog. *IEEE Transactions on Circuits and Systems for Video Technology*, 33:861–871, 2022.

[47] Xi Zhu, Zhendong Mao, Chunxiao Liu, Peng Zhang, Bin Wang, and Yongdong Zhang. Overcoming language priors with self-supervised learning for visual question answering. *arXiv preprint arXiv:2012.11528*, 2020.

Jin Liu is currently a Ph.D. candidate in the School of Control Science and Engineering at Shandong University, Jinan, China. He received his B.S. degree in the School of Control Science and Engineering at Shandong University, Jinan, China, in 2021. His research interests include cloud robots, robotic grasping, and deep learning.

Jialong Xie is currently working toward a Ph.D. degree in the School of Control Science and Engineering at Shandong University, Jinan, China. He received his B.Eng. degree in the School of Automation, Hangzhou Dianzi University, Hangzhou, China, in 2019. His research interests include cloud robots, human-robot interaction, and robot vision.

Fengyu Zhou received the B.E. and M.E. degrees from the Shandong University of Science and Technology, Jinan, China, in 1992 and 1996, and the Ph.D. degree in electrical engineering from Tianjin University, Tianjin, China, in 2008. He is currently a Professor at the School of Control Science and Engineering, Shandong University, Jinan. He has published more than 150 academic papers in international journals and conferences. His current research interests include evolutionary computation and cloud-intelligent robots. Prof. Zhou is the Chief Scientist of National Key Research and Development Projects, a Senior Member of the China Automation Society, a member of the Construction Robot Professional Committee, and the Director of the Science Popularization and Innovation Working Committee of the Shandong Automation Society.



Shengfeng He (Senior Member, IEEE) is an associate professor in the School of Computing and Information Systems, Singapore Management University. He was on the faculty of the South China University of Technology, from 2016 to 2022. He obtained B.Sc. and M.Sc. degrees from Macau University of Science and Technology in 2009 and 2011 respectively, and a Ph.D. degree from City University of Hong Kong in 2015. His research interests include computer vision and generative models. He is a senior member of IEEE and CCF. He serves as the lead guest editor of the IJCV, the associate editor of IEEE TNNLS, IEEE TCSVT, Visual Intelligence, and Neurocomputing. He also serves as the area chair/senior program committee of ICML, AAAI, IJCAI, and BMVC.