Test-Time Model Adaptation for Visual Question Answering With Debiased Self-Supervisions

Zhiquan Wen, Shuaicheng Niu, Ge Li, *Member, IEEE*, Qingyao Wu, Mingkui Tan, *Member, IEEE*, and Qi Wu

***Abstract*—Visual question answering (VQA) is a prevalent task in real-world, and plays an essential role in helping the blind understand the physical world. However, due to the real-world complexity, VQA test samples may come from a different distribution from the training data, resulting in unavoidable performance degradation. This similar issue also exists in the image recognition field, in which one most recent effective solutions is a test-time adaptation (TTA). TTA adapts a trained model at test time using only test samples, which provides a new idea to alleviate the analogous issue in VQA. However, naively introducing existing TTA methods (*e.g.,* test-time entropy minimisation) into VQA is imperfect and achieves only marginal performance gain. The reason is that prior methods do not consider the special nature of the VQA problem and ignore that 1) the biased samples in the dataset may have negative effects on test-time model adaptation, and 2) the model may have captured the biases in the dataset. In this paper, we propose Test-time Debiased Self-supervised (TDS) learning objectives for VQA model adaptation. Specifically, we minimise the entropy for those unbiased test samples. To identify these samples, we construct a negative sample for each test sample, and regard the test samples as unbiased if the output answers are different when feeding the test sample and the counterpart negative sample into the VQA model. Meanwhile, we also remove those samples with high prediction entropy from adaptation, making the test-time gradients more reliable. To hinder the model from excessively fitting the superficial correlations of the biased sample, we adopt the biased samples and the counterpart negative samples to assist the adaptation. Extensive experiments on the VQA-CP v1 and VQA-CP v2 datasets demonstrate the effectiveness of our TDS.**

***Index Terms*—Test-time adaptation, visual question answering, test-time debiased self-supervised.**

Manuscript received 23 November 2022; revised 8 March 2023 and 12 June 2023; accepted 29 June 2023. Date of publication 5 July 2023; date of current version 2 February 2024. This work was supported by the Ministry of Science and Technology Foundation Project 2020AAA0106900, National Natural Science Foundation of China (NSFC) 61836003 (key project), National Natural Science Foundation of China (NSFC) 62072190, CCF-Tencent Open Fund RAGR20220108, Program for Guangdong Introducing Innovative and Entrepreneurial Teams 2017ZT07X183. The Associate Editor coordinating the review of this manuscript and approving it for publication was Mrs. Si Liu. *(Zhiquan Wen and Shuaicheng Niu are with equal Contribution.) (Corresponding author: Mingkui Tan.)*

Zhiquan Wen, Shuaicheng Niu, Qingyao Wu, and Mingkui Tan are with the School of Software Engineering, South China University of Technology, Guangzhou, 510000, China (e-mail: sewenzhiquan@mail.scut.edu.cn; sensc@mail.scut.edu.cn; qyw@scut.edu.cn; mingkuitan@scut.edu.cn).

Ge Li is with the School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, Shenzhen 518055, China (e-mail: geli@ece.pku.edu.cn).

Qi Wu is with the School of Computer Science, University of Adelaide, Adelaide, SA 5005, Australia (e-mail: qi.wu01@adelaide.edu.au).

Digital Object Identifier 10.1109/TMM.2023.3292597

I. Introduction

VISION-AND-LANGUAGE tasks [1], [2], [3], [4], [5] become more and more pervasive in human daily life. For example, visual-language navigation (VLN) [3], [6], [7] tasks require agents to follow the indications from the humans. Visual question answering (VQA) [8], [9], [10] tasks can help visually impaired people to understand the physical world by question answering. The most advanced techniques in the above tasks often rely on deep neural networks (DNN) and have achieved outstanding performance. One assumption of the supervised DNN-based learning methods is the training and testing data distributions are similar, to guarantee that the deep model trained on training data can generalise to testing data well. However, the test scenarios in the real world may be more complex, and thus the test samples may be drawn from a different distribution from the training data. For example, the answer distribution of training and testing data in the VQA-CP v2 [11] dataset is different, resulting in severe performance degradation of prior methods [11], [12]. Thus, when encountering distribution shifts of test samples, appropriately adapting the deployed model is necessary to guarantee performance.

Recently, in the image classification field, fully test-time adaptation approaches [14], [15], [16] have been proposed to address the analogous distribution shift issues. These methods are able to online adapt any pre-trained model on a mini-batch or single test sample, and can significantly improve the model performance on out-of-distribution (OOD) test datasets. Tent [14] first proposes to adapt a pre-trained model at test time by minimising the entropy of test samples, which significantly boosts the model performance. However, as pointed out by EATA [15], not all test samples can benefit test-time model adaptation. Thus, the authors identify unreliable samples and omit the gradients produced by these samples, which further improves the adaptation performance and efficiency. In addition to entropy minimisation, MEMO [16] further introduces prediction consistency maximisation of one sample's different augmentation views, which enables the adaptation of even one single test sample.

Inspired by the success of the above fully test-time adaptation methods on image classification, we seek to alleviate the distribution shift issue in VQA through the test-time adaptation manner. However, naively transferring prior methods into the VQA field may be unsuitable. Specifically, applying previous TTA methods to the VQA field still poses the following challenges: 1) These methods ignore the bias issue [12], [17] in the

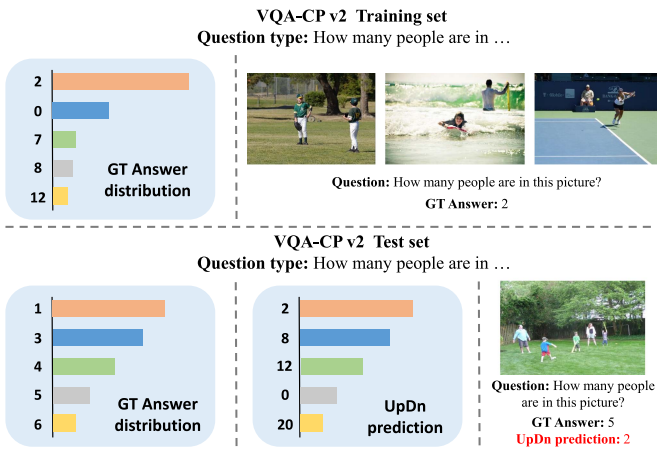

Fig. 1. Examples of the answer distribution about the question "How many people are in . . . " on the VQA-CP v2 [11] datasets. Specifically, on the training set, the answers are mainly distributed in "2", while on the test set, "2" only occupies little part of the answers. UpDn [13] model captures the biases between the questions and answers in the training set, and thus tends to adopt the "bias" to answer the question instead of the reasoning ability.

process of test-time adaptation. Since most machine learning datasets inevitably contain biases (For example, in Fig. 1, the VQA-CP v2 dataset has severe bias issue), minimising the entropy of biased samples to update the model may have adverse effects on the model performance. 2) These methods ignore that the model may have captured the biases in the datasets (As shown in Fig. 1, Updn [13] model has captured the biases in the training set). Alleviating the bias issue during the test-time adaptation is important, which can promote the model to use reasoning ability to answer questions. However, how to alleviate the bias at test time is still an open question.

To address these challenges, in this paper, we propose a novel method named test-time debiased self-supervisions (TDS) to improve the model performance on the OOD dataset at test time. Specifically, to eliminate the entropy of biased samples, we first require to recognise the biased samples. The bias issue usually denotes that the VQA model captures the shortcut between the questions and answers to answer the questions, without considering the image. From this perspective, the VQA model may output the same answer given biased sample (v_i, q_i) and its counterpart negative sample $(\hat{v}_i, q_i)$. Thus, to find the biased samples, we first consider constructing the negative samples by randomly sampling one image $\hat{v}_i$ in a mini-batch data for each sample. Then regarding the samples as biased samples if the output answers are the same when feeding the samples and their counterpart negative samples to the VQA model. Moreover, we set a threshold to filter the samples with high entropy before the model adaptation to improve adaptation stability. Finally, to alleviate the bias at test time, we consider inhibiting the model to excessively fit the superficial correlations of the biased samples, which can be achieved by minimising the possibility of predicting the answer of biased samples and their counterpart negative samples. Experiments on the VQA-CP v1 and VQA-CP v2 datasets demonstrate the effectiveness of our proposed TDS.

Our contributions can be summarised as follows:

- We investigate the effects of test-time adaptation in VQA, and provide a new perspective for the VQA field to boost a pre-trained model's performance when test samples have a different distribution from the training data.
- We devise a debiased test-time entropy minimisation objective that identifies biased samples and then omits the gradients of these samples during test-time adaptation.
- We propose to alleviate the bias in the VQA model at test time by minimising the prediction of the highest score of the biased samples and the counterpart negative samples.

II. Related Work

A. Visual Question Answering

Visual question answering (VQA) task requires an agent to answer the textual question based on a corresponding image, which has been proposed by [18]. This task is challenging, and requires the agent to master a powerful reasoning ability to be accomplished. Nevertheless, many researchers have make many efforts to accomplish this task, which can be categorised mainly into three types, namely attention mechanism based methods [19], [20], [21], graph learning based methods [1], [22], [23], and knowledge based methods [24], [25], [26]. Moreover, to evaluate these methods, many benchmark datasets have been proposed, such as VQA v1 [18] and VQA v2 [27]. However, existing datasets inevitably have biases, which hinder the development of the VQA. For example, on the VQA v1 [18] dataset, one can answer "yes" to the question type "Do you see a . . . " without considering the rest part of the question and the information of the image, which would obtain around 87% accuracy. To alleviate this issue, Goyal et al. [27] balance the VQA v1 dataset by collecting complementary images to form the VQA v2 dataset. However, these datasets still contain superficial correlations that are easy captured by the VQA models. To further evaluate the true reasoning ability of the VQA model, Agrawal et al. constructed the out-of-distribution (OOD) dataset named VQA-CP v2 [11] by re-organising the training and validation sets of the VQA v2 dataset, where the answer distributions in training and test sets of VQA-CP v2 are different. Thus, the VQA-CP v2 dataset provides a new testbed for verifying the true reasoning ability and robustness of the VQA models.

B. Overcome Biases in VQA

Although existing methods [19], [22], [28] can achieve promising performance on the in-domain datasets (e.g., VQA v2 [27] dataset), the performance would decrease severely when encountering out-of-domain datasets (e.g., VQA-CP v2 dataset [11]), which limits the deployment of the VQA model in real-world. To alleviate the bias issue in VQA, some methods [12], [29], [30], [31] seek to introduce debiased techniques in training time. Specifically, these methods can be divided into two types based on whether adopting data augmentation techniques. The methods without data augmentation seek to directly weaken the language bias [12], [17], [30] or augment the visual

representation [32], [33], while data augmentation based methods [29], [34], [35], [36], [37] attempt to balance the biased training set to implicitly alleviate the bias issue.

For the methods without data augmentation, some methods [12], [17], [30], [36] constructed additional unimodal branches in training time to capture the biases to be removed, which introduces additional computational overhead. Moreover, CF-VQA [30] even introduces additional additional parameters in the inference phase. To make the VQA models focus more on the visual or text information, some methods [32], [33] introduce human based annotations to assist the training process to strengthen the visual grounding. However, these methods require expensive human based annotations that are hard to be obtained.

For the data augmentation based methods, CSS-VQA [29] and Mutant [37] methods generated massive counterfactual samples by masking the vital objects and words in the images and questions, respectively, to assist the training process. However, these methods required expensive annotations. Moreover, to mitigate the dependency on the expensive annotations, some methods [34], [35], [36], [38] constructed negative samples based on the available samples to balance the dataset.

These methods address the bias issue by introducing debiased techniques mainly in training time. In this paper, we argue that the bias issue is also inevitable in test-time adaptation, and provide an additional perspective to alleviate the bias in the fully test-time adaptation setting. Moreover, these methods usually adopt the test set to select the best model. However, in real-world applications, the information about the OOD distribution should also be unavailable until we evaluate the model. In this sense, when one deployed model encounters data that have different distributions from the training data, alleviating the dataset shift in test time based on the test data may be a better solution.

C. Test-Time Adaption

Test-time adaptation [14], [15], [39], [40], [41], [42], [43], [44], [45], [46] seeks to adapt a model learned from a source training domain to a potentially shifted target test domain, where only test samples are used for adaptation. Specifically, TTT first proposes the pipeline of test-time training [39], in which the authors train a classification model using both a supervised learning objective and self-supervision (rotation prediction [47]). Then, given a test sample, TTT will first train the model using this sample with self-supervised rotation prediction and then use the updated model for the final prediction. Although effective at handling test shifts, TTT alters the model training process, which may be very impractical in some real-world applications. To avoid altering the training process and relying on the access to original training data, Tent [14] proposes a fully test-time adaptation, in which the authors put forward to minimise the unsupervised entropy of test samples. Based on the Tent, MEMO [16] adopts the data augmentation techniques to augment the images, and updates the model with these samples. Moreover, DDA [48] proposed to first adopt the diffusion model to mitigate the shift of the image, and then predict the processed samples without updating the model. However, during the test-time adaptation process, compared with Tent, MEMO and DDA are computationally expensive and inefficient. Similar to MEMO, TPT [42] proposes to adapt the prompt parameters guided by the entropy of the samples to improve the CLIP [49] model's performance. Moreover, CoTTA [43] considers a new continual test-time adaptation setting, *i.e.,* target domain distribution can change over time, and devises a novel method to address the error accumulation and catastrophic forgetting issues in the setting. Motivated by the inherent uncertainty around the conditions that will ultimately be encountered, LAME [44] introduces the Laplacian Adjusted Maximum-likelihood Estimation objective, in which the objective is addressed by adapting the model's output in the concave-convex procedure.

After that, the idea of test-time adaptation has been applied to many real-world applications, such as reinforcement learning [50], [51], human pose estimation [52], single image dehazing [53], dynamic scene deblurring [54], and so on. In our work, we consider the special nature of the visual question answering (VQA) problem and then devise an efficient fully test-time adaptation method for VQA.

III. Problem Definition

In visual question answering (VQA), many models obtain promising performance in the IID dataset (e.g., VQA v2 dataset [27]) while suffering severe performance degradation when the answer distribution differs between training and testing (e.g., VQA-CP dataset [11]). This phenomenon hinders the usage when the deployed VQA model encounters data from different answer distributions. To alleviate this issue, in this paper, we seek to study how to adapt a pre-trained VQA model at test time to improve the model performance given only the pre-trained VQA model and the test data, without the test label and source training data.

Without loss generality, let M_{θ_p} be a VQA model that is pre-trained on labelled training data $\mathcal{D}_s = \{(v_i, q_i, a_i)\}_{i=1}^{N_s}$, where θ_p denotes the parameters of the pre-trained VQA model. When a VQA model is deployed for the first time, in some cases, we may not have access to the source data due to privacy reasons. And there are no ground truth labels provided because they are test data ($\mathcal{D}_t = \{(v_j, q_j)\}_{j=1}^{N_t}$). In this way, our aim is to boost the model performance on the out-of-distribution (OOD) test set with only the test data, which can be formulated as:

$$\min_{\theta_p} -\frac{1}{N_t}\sum_{j=1}^{N_t}\mathcal{L}((v_j, q_j); \theta_p), \tag{1}$$

where $(v_j, q_j) \in \mathcal{D}_t$, and θ_p are the parameters of the VQA model that requires to be updated. However, how to adapt the VQA model to the OOD test set is still an open question.

IV. Test-Time VQA Model Adaptation With Debiased Self-Supervisions

In this paper, we seek to improve the VQA model performance on the out-of-distribution (OOD) test set at test time by using only the unlabelled test samples, which is vital in real-world model deployment. To this end, we propose a TDS method.

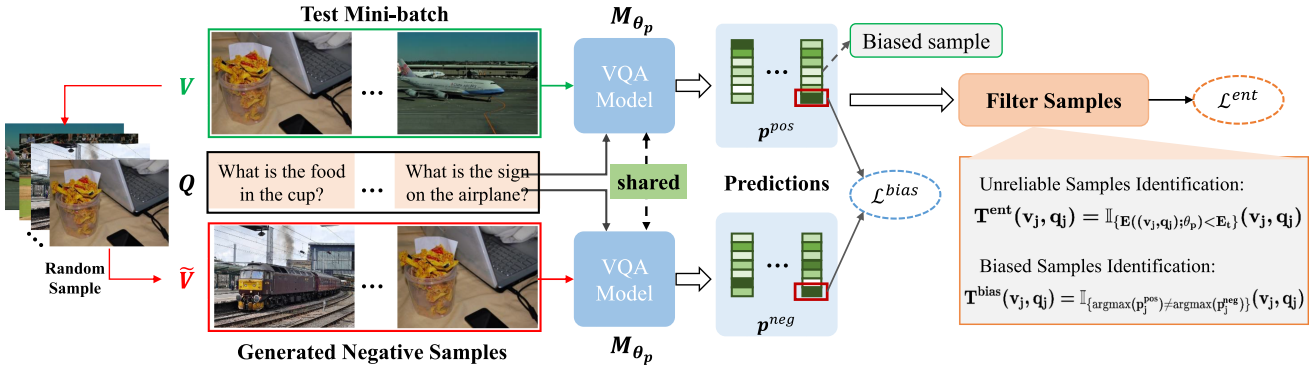

Fig. 2. Overview of our TDS method. Given a pre-trained model M_{θ_p}, we seek to conduct test-time adaptation to improve the performance by updating the parameters θ_p with only the test data without a label. Specifically, given the predictions $\mathbf{p}^{pos}$, we consider filtering the samples with high entropy to mitigate the negative effect of unreliable samples. Moreover, we construct the negative samples by randomly sampling one image in a mini-batch for each sample to assist to find and filter the biased samples. Finally, we adopt the biased samples and the counterpart negative samples to alleviate the bias in the VQA model.

Specifically, to adapt the pre-trained VQA model M_{θ_p} to the OOD test set, one intuitive way is to regard the learning objective $\mathcal{L}(\cdot)$ as an entropy minimisation problem, *i.e.,*

$$\min_{\theta_p} \frac{1}{N_t}\sum_{j=1}^{N_t} E((v_j, q_j); \theta_p)$$
$$= \min_{\theta_p} -\frac{1}{N_t}\sum_{j=1}^{N_t}\sum_{a\in\mathcal{A}} M_{\theta_p}(a|v_j, q_j)\log M_{\theta_p}(a|v_j, q_j), \quad (2)$$

where $\mathcal{A}$ denotes the collection of the candidate answers. This paradigm has been studied in the test-time adaption on the image classification task [14], [15]. However, as is known to all, most machine learning datasets inevitably contain biases. The VQA model may capture the biases instead of learning the reasoning ability. In this sense, we should filter the biased sample before adopting the entropy of the samples to adapt the model (Section IV-A1). Moreover, not all entropy of the samples for updating the model is helpful [15], we thus further filter the samples with high predicted entropy (Section IV-A2). Besides filtering the entropy of the biased sample, we further propose to inhibit the VQA model to learn the biases from the biased sample at test time (Section IV-B). By adopting the above three techniques, the VQA model is able to adapt to the OOD test set at test time well.

A. Selective Test-Time Entropy Minimisation

1) Biased Samples Identification: Due to the existence of biases in the dataset, the VQA model inevitably captures the biases, *i.e.,* the model excessively relies on superficial correlations to answer questions, instead of mastering the reasoning ability [17], [30], [36]. In this sense, adapting the VQA model using biased samples may be detrimental.

To alleviate the above issue, we require to identify the biased samples and then remove the entropy generated from these samples. Thus, the issue is transformed into how to recognise the biased samples. In general, the bias issue usually denotes the superficial correlations between the questions and answers [17], [30]. In other words, the model tends to answer the questions without focusing on the images. From this perspective, given a pair of test sample (v_j, q_j), when answering the question q_j, if the sample is biased, the VQA model would output similar predictions regardless of which image it is.

Based on this intuition, we consider constructing the negative samples as the counterpart of the test sample to recognise the biased samples. Specifically, given a mini-batch of test data $\{(v_j, q_j)\}_{j=1}^{B}$, for each sample (v_j, q_j), we construct the negative sample $(\tilde{v}_j, q_j)$ by randomly sampling one image $\tilde{v}_j$ from the mini-batch of images $\{v_j\}_{j=1}^{B}$. By feeding the test samples and the counterpart negative samples to the VQA model, we would obtain two types of predictions $\mathbf{p}^{pos}$ and $\mathbf{p}^{neg}$, *i.e.,*

$$\mathbf{p}_j^{pos} = M_{\theta_p}(\mathcal{A}|v_j, q_j),\ \mathbf{p}_j^{neg} = M_{\theta_p}(\mathcal{A}|\tilde{v}_j, q_j). \quad (3)$$

Then we regard the samples as the biased samples when the top-1 answer between the two predictions are the same, *i.e.,* $\text{argmax}(\mathbf{p}_j^{pos}) == \text{argmax}(\mathbf{p}_j^{neg})$. Thus, the process of filtering the biased samples can be formulated as:

$$\mathbf{T}^{bias}(v_j, q_j) = \mathbb{I}_{\{\text{argmax}(\mathbf{p}_j^{pos})\neq\text{argmax}(\mathbf{p}_j^{neg})\}}(v_j, q_j), \quad (4)$$

where the operation argmax is to find an index that has a maximum value among the predictions, $\mathbb{I}(\cdot)$ is an indicator function, *i.e.,* the value is 1 if $\text{argmax}(\mathbf{p}_j^{\text{pos}})$ is not equal to $\text{argmax}(\mathbf{p}_j^{\text{neg}})$, and 0 otherwise. In this way, we are able to alleviate the negative effect of the biased samples when adapting the VQA model to the OOD test set.

Note that when test batch size B equals 1, we construct negative image $\tilde{v}_j$ for a single test sample by generating the Gaussian noises of the same size as the image features. Results in Section V-D2 show that this strategy works well.

2) Unreliable Samples Identification: The samples with high entropy demonstrate the predictions of the VQA model on these samples are uncertain. Thus, the gradient generated by the entropy loss of these samples may be unreliable [15]. To get rid of the negative effect of these samples, we introduce a pre-defined entropy threshold E_t to filter the samples with high entropy, which can be formulated as:

$$\mathbf{T}^{ent}(v_j, q_j) = \mathbb{I}_{\{E((v_j,q_j);\theta_p)<E_t\}}(v_j, q_j), \quad (5)$$

where $E((v_j, q_j); \theta_p)$ denotes the entropy of the sample (v_j, q_j) based on the prediction of VQA model M_{θ_p}, and $\mathbb{I}(\cdot)$ is an indicator function, *i.e.,* the value is 1 if the entropy of sample j smaller than the threshold E_t, and 0 otherwise. In this way, we are able to filter the samples with high entropy and remain relatively reliable samples to update the VQA model at test time.

Note that the threshold E_t can be defined based on the number of candidate answers C, *i.e.,* $E_t = \alpha \ln C$, where $\alpha \in [0, 1]$ is a hyper-parameter that adjusts the threshold of the entropy. More results about the ablation studies of the α can be found in Section V-D1.

3) Selective Entropy Loss: Based on the above, when adapting the VQA model at test time, we remove the entropy of unreliable and biased samples to mitigate the negative effect of these samples. Based on the (4) and (5), the selective entropy loss $\mathcal{L}^{ent}$ is formulated as:

$$\mathcal{L}^{ent} = \frac{1}{N_t} \sum_{j=1}^{N_t} \mathbf{T}^{ent}(v_j, q_j) \mathbf{T}^{bias}(v_j, q_j) E((v_j, q_j); \theta_p). \quad (6)$$

B. Instance-Level Test-Time Bias Alleviation

The above techniques mainly filter the samples that have high entropy or are biased when adapting the VQA model to the OOD test set. Different from above, in this part, we seek to alleviate the bias directly in the VQA model.

Before introducing the adapting technique, we first require to recognise the true biased samples. Although we have found the biased samples in Section IV-A1, some biased samples may be pseudo due to the uncertainty of the model to the samples. In other words, when encountering samples with high entropy, feeding the counterpart negative samples to the VQA model may obtain similar output due to the uncertainty of the model, leading to the samples being regarded as biased samples. Thus, the biased samples may contain two characterises, namely, 1) the samples should be reliable, *i.e.,* the samples have low entropy; 2) Feeding the VQA model with the initial samples and counterpart negative samples, respectively, would obtain the same answers. In this way, we obtain the true biased samples by $\tilde{\mathbf{T}}^{bias}(v_j, q_j) = \mathbf{T}^{ent}(v_j, q_j)(1 - \mathbf{T}^{bias}(v_j, q_j))$.

Generally speaking, the bias issue is that the VQA model excessively fits the superficial correlations between the questions and answers. To alleviate the bias, inspired by [34], [35], [36], when feeding the biased test samples and counterpart negative samples to the VQA model, we seek to hinder the VQA model from correctly answering the question, which can be achieved by minimising the prediction of the ground-truth answer. However, since we cannot have the access to the ground-truth answer of the test samples, we consider the prediction with the highest score as the pseudo label, *i.e.,* $k = \text{argmax}(\mathbf{p}_j^{pos})$. In this way, the debiased loss $\mathcal{L}^{d\_bias}$ can be formulated as:

$$\begin{aligned}\mathcal{L}^{d\_bias} = &\tilde{\mathbf{T}}^{bias}(v_j, q_j)\text{softmax}(\mathbf{p}_j^{pos})[k] \\ &+ \tilde{\mathbf{T}}^{bias}(v_j, q_j)\text{softmax}(\mathbf{p}_j^{neg})[k]. \quad (7)\end{aligned}$$

Note that if k corresponds to the ground-truth answer, optimising the VQA model with the first part of $\mathcal{L}^{d\_bias}$ is able to alleviate the bias (*i.e.,* do not excessively fit the superficial correlations), to some extent. Minimising the last part of $\mathcal{L}^{d\_bias}$ is able to promote the VQA model to focus on the image, since the model cannot answer the question correctly when given the mismatched question and image.

In total, our method mainly contains two types of loss, namely, the selective entropy loss $\mathcal{L}^{ent}$ and debiased loss $\mathcal{L}^{bias}$, which can be formulated as:

$$\mathcal{L}^{total} = \gamma_1 \mathcal{L}^{ent} + \gamma_2 \mathcal{L}^{d\_bias}. \quad (8)$$

We set $\gamma_1 = \gamma_2 = 1.0$ without any tuning for all experiments, showing the metric of hyper-parameters insensitivity.

Algorithm 1: The Pipeline of Proposed TDS.

Require: Test samples $\mathcal{D}_t = \{(v_j, q_j)\}_{j=1}^{N_t}$, the pre-trained model $M_{\theta_p}(\cdot)$, batch size B.

1: **for** a mini-batch $\mathcal{D}_b = \{(v_b, q_b)\}_{b=1}^{B}$ in $\mathcal{D}_t$ **do**
2: Calculate the predictions $\mathbf{p}^{pos}$ for the samples $(v_b, q_b) \in \mathcal{D}_b$ with the model $M_{\theta_p}(\cdot)$ via (3).
3: Obtain the indicators $\mathbf{T}^{ent}(v_j, q_j)$, $\mathbf{T}^{bias}(v_j, q_j)$ and $\tilde{\mathbf{T}}^{bias}(v_j, q_j)$ via (4) and (5).
4: Update the model M_{θ_p} with (8).
5: **end for**

Ensure: The predictions $\{\mathbf{p}^{pos}\}_{j=1}^{N_t}$ for all $(v_j, q_j) \in \mathcal{D}_t$.

V. Experiments

In this section, we evaluate the proposed method TDS in the setting of fully test-time adaptation on the out-of-distribution (OOD) benchmark datasets VQA-CP v1 [11] and VQA-CP v2 [11], where the training and testing answer distributions of these datasets are different.

A. Datasets and Compared Methods

1) VQA-CP v2 Dataset: To evaluate the robustness of the VQA model, Agrawal et al. [11] re-organised the training and validation splits of the VQA v2 [27] dataset, in which the answer distributions between the training and test sets of the VQA-CP v2 dataset are highly different. Specifically, the training set of VQA-CP v2 contains approximately 121 k images and 438 k questions, while the test set contains approximately 98 k images and 220 k questions. Training the traditional VQA model [13], [55], [56] based on this dataset would inevitably capture the biases, and achieve poor performance.

2) VQA-CP v1 Dataset: Similar to the VQA-CP v2 dataset, the VQA-CP v1 dataset is also constructed by re-splitting the training and validation sets of the VQA v1 [18] dataset to make the answer distributions between the training and test splits are different. Specifically, VQA-CP v1 dataset contains approximately 118 k images and 244 K questions for training, and contains about 87 k images and 125 K questions for testing.

3) Compared Methods: We compare our TDS with the state-of-the-art fully test-time adaptation methods, including Tent [14], ETA [15], TPT [42], CoTTA [43], and LAME [44]. Specifically, Tent minimises the entropy of the samples at test

TABLE I
COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON THE VQA-CP V2 AND VQA-CP V1 TEST SET IN TERMS OF ACCURACY (%)

	Model	VQA-CP v2 test set				VQA-CP v1 test set			
		All	Yes/No	Number	Other	All	Yes/No	Number	Other
1	UpDn [13]	40.94	43.87	12.90	47.09	38.78	42.78	13.38	45.01
2	Tent [14]	41.17	43.88	13.95	47.22	38.79	42.77	13.45	45.02
3	ETA [15]	41.28	43.81	13.76	47.51	38.95	42.89	13.73	45.15
4	TPT [43]	41.15	43.85	14.16	47.13	38.64	42.39	13.54	45.00
5	CoTTA [44]	41.01	43.86	12.97	47.21	38.80	42.74	13.39	45.09
6	LAME [45]	41.37	42.91	**15.70**	47.61	38.46	41.59	**13.84**	45.24
4	TDS (ENT)	41.27	43.80	14.28	47.36	38.71	42.52	13.49	45.06
5	TDS (ENT + BIAS)	45.32	**64.11**	11.16	44.84	43.68	54.64	13.30	44.99
6	TDS (ENT + BIAS + BIAS SAMPL)	46.12	62.11	10.50	47.51	45.44	58.09	13.70	45.60
7	**TDS (ALL)**	**46.33**	62.55	10.53	**47.66**	**45.55**	**58.22**	13.82	**45.70**

The baseline model is UpDn. 'ENT' denotes removing the sample with high entropy Via eq. (5). 'Bias' denotes removing the sample with bias Via eq. (4). 'BIAS SAMPL' denotes minimising the score with the highest prediction on the true bias samples by part of eq.(7).

time to adapt the models. Inspired by Tent, ETA still adopts the entropy of the samples as the optimisation objective to adapt the models, in which the entropy generated by unreliable and redundant samples is excluded. TPT [42] adapts the model parameters guided by the entropy of the augmented samples to improve model performance. CoTTA [43] uses the weight-averaged and augmentation-averaged predictions as pseudo-labels to guide the test-time adaptation. LAME [44] introduces the Laplacian Adjusted Maximum-likelihood Estimation objective, in which the objective is addressed by adapting the model's output in the concave-convex procedure.

B. Implementation Details

In a real-world scenario, the deployed VQA system would first output an answer immediately when the users pose a question, and then collect the question-image-answer pairs for the subsequent model adaptation. To simulate this scenario, we suppose a VQA model pre-trained on the VQA-CP [11] v2 training set as a deployed VQA system, and the VQA-CP v2 test set is regarded as the questions that the users proposed. In this way, the test-time adaptation process in VQA is step-wise, *i.e.,* for each mini-batch data, the VQA system predicts the answer first, and then updates the VQA model based on the current data. The detailed pipeline of our TDS can be found in Algorithm 1 of the manuscript.

On the VQA-CP v1 and VQA-CP v2 dataset, following existing debiased VQA methods [34], [36], we extract the object features of the images by adopting the object detection method Faster-RCNN [57] pre-trained by [13]. In each image, we extract the top-36 object features, and the dimension of each feature is 2048. Moreover, we truncate or pad each question into the same length, *i.e.,* 14, and then encode each word with Glove [58] embedding with the dimension of 300. The dimension of the hidden state is set to 512. On the VQA-CP v2 dataset, we collect 2274 candidate answers that are occur more than 9 times in the training set, while we obtain 1691 candidate answers on the VQA-CP v1 dataset. We implement our entire method based on PyTorch [59].

Note that our method is model-agnostic, which can be applied to different baseline models. In our experiments, we select three mainstream baseline models, namely, UpDn [13], ViLBERT [56], and LXMERT [55].

We adopt the SGD optimiser with momentum is 0.9 to update the models. The learning rate of the UpDn model is set to 0.01, and the batch size is set to 512. Due to the limitation of the memory, in ViLBERT, the batch size is set to 32 and the learning rate is 0.0002, while in LXMERT the batch size is set to 32, and the learning rate is 0.0001. The source code and the pre-trained models are available at TDS.

C. Evaluation on the VQA-CP v1 and VQA-CP v2 Datasets

1) Quantitative Results: We compare our TDS with other state-of-the-art methods on the VQA-CP v1 and VQA-CP v2 datasets [11] in terms of Accuracy, and report the experimental results in Table I. From these results, we have the following observations: 1) The compared methods (e.g., Tent [14], ETA [15], CoTTA [43], and TPT [42]) obtain higher accuracy compared with the baseline model UpDn. These results demonstrate that adopting entropy minimisation as the learning objective is able to improve model performance in the setting of fully test-time adaptation in VQA, to some extent. Moreover, LAME introduces Laplacian Adjusted Maximum-likelihood Estimation objective to adjust the model's output, which also improves the model performance. However, these methods ignore the bias issue, and thus the improvement is marginal. 2) Our TDS outperforms all the compared methods. Specifically, our TDS surpasses UpDn [13], Tent [14], ETA [15], LAME [44], CoTTA [43], and TPT [42] by around 5% on the VQA-CP v2 dataset, while exceeding all the compared methods (*i.e.,* UpDn, Tent, ETA, LAME, CoTTA, and TPT) by around 6.5% on the VQA-CP v1 dataset. These results further demonstrate that in the VQA field, only considering removing the samples with high entropy (*i.e.,* ETA) is not enough in the fully test-time adaptation setting, while alleviating the bias issue is also important and necessary.

Moreover, on the VQA-CP v2 dataset, our TDS performs worse than other methods on the answer type of 'Number'. To analyse the results, we visualise the answer distributions on the question type 'How many . . . ' in Fig. 3. From these results, ETA and Tent methods ignore the bias issue, and excessively fit

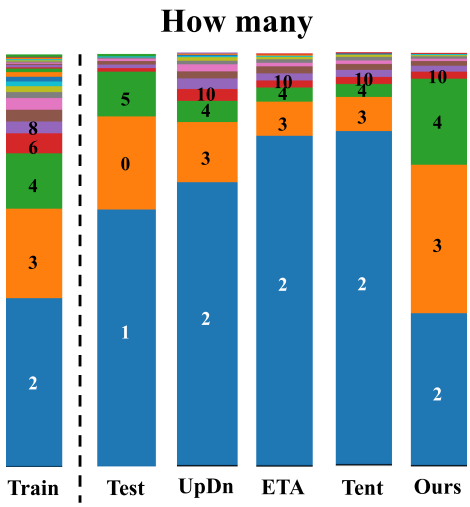

Fig. 3. Qualitative comparisons of the answer distributions about the question type 'How many . . . ' among Updn [13], Tent [14], ETA [15], and our TDS on the VQA-CP v2 test set.

the biased samples more than the UpDn model, thus achieving comparable results with UpDn. Nevertheless, our TDS considers the bias issue, and thus the answer distribution is different from that in the baseline model UpDn. However, since missing the supervision from the ground-truth label, our TDS performs unsatisfactorily on the answer type 'Number'.

2) Qualitative Results: To further demonstrate the effectiveness of our proposed methods, we provide some qualitative results on the VQA-CP v2 test set in Figs. 4, and 5. As shown in the results in Fig. 4, the Tent and ETA methods can not find the target object in the image, and thus output the wrong answers. Instead, our method is able to locate the target object in the image with high weight, and make a correct prediction. Moreover, in Fig. 5, we visualise the answer distributions of different methods about different question types (*i.e.,* "Dose the . . . ", "What color . . . ", and "Is this . . . "). From these results, our TDS is able to adjust the biased answer distributions of the UpDn model trained from the training set to approach the test answer distribution. In contrast, other compared methods still suffer from the bias issue and obtain similar answer distribution with the training set but different from the test set. These qualitative results further demonstrate the effectiveness of our TDS.

D. Ablation Studies

1) Effect of the Hyper-Parameter α That Adjusts the Threshold of the Entropy: As referred to in Section IV-A2, we require to determine the pre-defined entropy threshold E_t. Specifically, in the VQA-CP v2 dataset, we collect 2274 candidate answers that occur more than 9 times in the training set, then the maximum entropy is $\ln 2274 \approx 7.73$. Based on the maximum entropy, we conduct ablation studies on the hyper-parameter α that can adjust the threshold of the entropy, and show the experimental results in Table II. From these results, we find that with the increase of the α, the accuracy of our TDS decreases gradually, which further verifies that the samples with high entropy would hurt

TABLE II
EFFECT OF THE α (REFER TO IN SECTION IV-A2) THAT ADJUSTS THE THRESHOLD OF THE ENTROPY ON THE MODEL PERFORMANCE IN TERMS OF ACCURACY (%). THE BASELINE MODEL IS UPDN

Model	α	VQA-CP v2 test set			
		All	Yes/No	Number	Other
TDS (ENT)	0.1	41.25	43.68	**14.30**	47.36
	0.2	**41.27**	43.80	14.28	**47.36**
	0.4	41.21	43.84	14.02	47.29
	0.6	41.19	43.87	13.99	47.25
	0.8	41.17	**43.89**	13.94	47.22
	1.0	41.17	**43.89**	13.94	47.22

TABLE III
EFFECT OF BATCH SIZE (BS) ON THE MODEL PERFORMANCE IN TERMS OF ACCURACY (%). THE BASELINE MODEL IS UPDN

Model	BS	VQA-CP v2 test set			
		All	Yes/No	Number	Other
TDS	1	45.58	58.06	14.72	47.51
	2	**48.28**	66.23	**16.74**	47.53
	4	48.03	**66.87**	10.00	**48.60**
	8	47.22	64.29	10.69	48.29
	16	47.09	64.26	10.61	48.10
	32	47.00	63.75	10.48	48.25
	64	46.96	63.96	10.51	48.04
	128	46.81	63.41	10.50	48.07
	256	46.51	61.94	10.72	48.28
	512	46.33	62.55	10.53	47.66
	1,024	46.13	60.72	10.70	48.21

the model performance in the setting of fully test-time adaptation. Moreover, our TDS (ENT) method with $\alpha = 0.1$ achieves comparable but slightly worse than that with $\alpha = 0.2$. The small α corresponds to the strict threshold E_t, which may exclude the samples that are helpful for the test-time model adaptation in VQA.

2) Effect of the Batch Size on the Model Performance: To evaluate the batch size on the model performance, we conduct experiments on different batch size, and present the experimental results in Table III. From the results, we have the following observations: 1) With the decrease in the batch size, the overall accuracy improves gradually. In this paper, we do not carefully tune the batch size and simply set it to a universal one (*i.e.,* batch size is 512), which is the same as that in the training process. Moreover, small batch size denotes the VQA models are adapted more frequently, and thus further improve the VQA models' performance, which is reasonable and practical. 2) Our TDS method is robust on arbitrary batch size, which demonstrates the superiority of our TDS. 3) Following ETA [15] and Tent [14], we also apply our TDS to an extreme condition, *i.e.,* batch size is 1. From the results, our TDS with batch size 1 achieves comparable performance compared to that equipped with any batch size, embodying the robustness of our method, which demonstrates our TDS is able to apply to real-world scenarios.

3) Effect of Each Component on the Model Performance: To demonstrate the effectiveness of each component in our proposed method, we conduct ablation studies regarding each component in our TDS on the VQA-CP v1 and VQA-CP v2

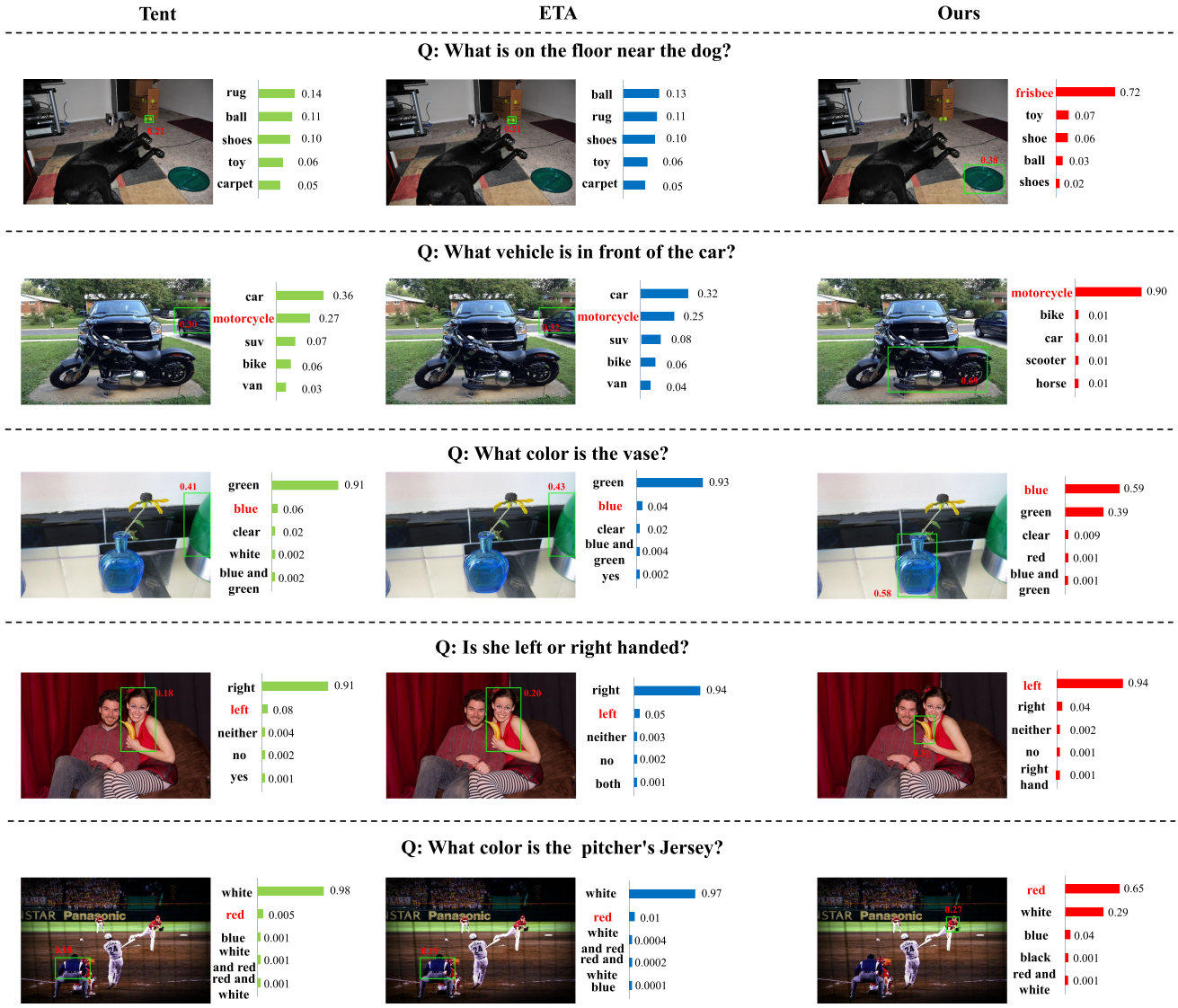

Fig. 4. Qualitative comparisons among Tent [14], ETA [15], and our TDS on the VQA-CP v2 test set. For each example, we put the bounding box with the highest attention weights in the image and show the answers with the top-5 predictions. The bold, red answer is the ground-truth answer.

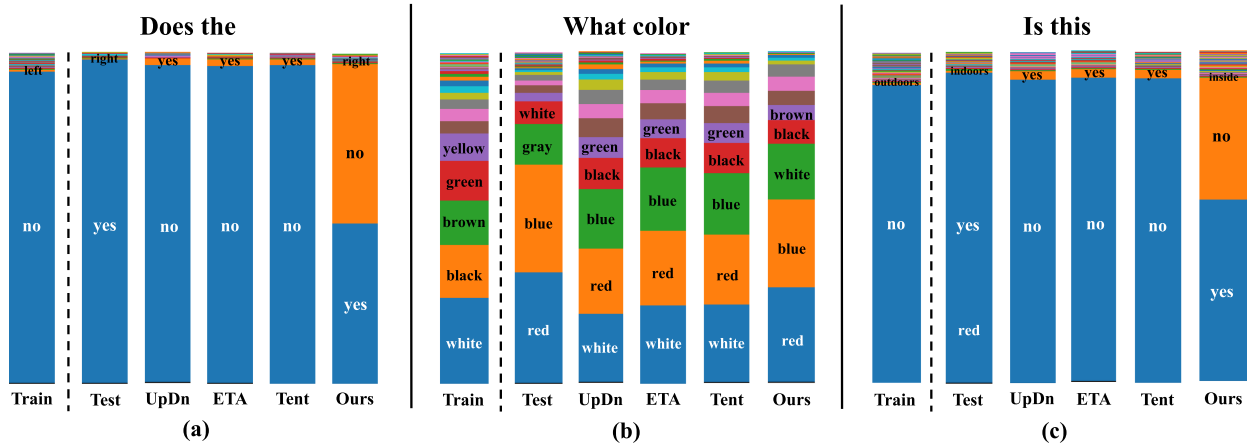

Fig. 5. Qualitative comparisons of the answer distributions about different question types among UpDn [13], Tent [14], ETA [15], and our TDS on the VQA-CP v2 test set.

TABLE IV
EXPERIMENTAL RESULTS OF OUR TDS WITH DIFFERENT BACKBONE MODELS ON THE VQA-CP V1 AND VQA-CP V2 DATASET. WE OBTAIN ALL BASELINE MODELS BASED ON THEIR OFFICIAL GITHUB REPOSITORY

Model	VQA-CP v2 test set				VQA-CP v1 test set			
	All	Yes/No	Number	Other	All	Yes/No	Number	Other
ViLBERT [57]	40.75	43.87	14.94	46.20	39.59	45.07	**15.14**	43.97
ViLBERT + TDS	**43.59**	**50.49**	**16.39**	**47.43**	**43.27**	**53.55**	14.83	**44.46**
LXMERT [56]	41.72	43.61	**14.17**	48.30	38.21	40.77	14.81	45.06
LXMERT + TDS	**45.07**	**54.24**	12.23	**49.27**	**42.02**	**49.67**	**15.40**	**45.11**
UpDn [13]	40.94	43.87	**12.90**	47.09	38.78	42.78	13.38	45.01
UpDn + TDS	**46.33**	**62.55**	10.53	**47.66**	**45.55**	**58.22**	**13.82**	**45.70**

datasets, and the experimental results are shown in Table I. From these results, we have the following observations: 1) Filtering the samples with high entropy only is able to promote the model performance in fully test-time adaptation (refer to Lines 1–4), which demonstrates that the samples with high entropy have a negative effect on adapting the model. 2) When further filtering the biased samples, the model performance would be improved again (*i.e.,* from 41.27% to 45.32% on the VQA-CP v2 dataset, while from 38.71% to 43.68% on the VQA-CP v1 dataset). These results demonstrate that alleviating the bias issue in the TTA setting is significant and not ignored. 3) Except for filtering the biased samples, alleviating the bias issue using the biased samples would improve the model performance (*i.e.,* refer to Lines 5–7). The results show that alleviating the bias in the biased VQA model has a positive effect on the model adaptation. In total, these results demonstrate the effectiveness of each component in our method on the model performance.

4) Evaluation of Trainable Parameters: To further verify the effect of the trainable parameters, we conduct experiments regarding the number of trainable parameters on the VQA-CP v2 test set with the baseline model UpDn. Specifically, we consider an additional variant that trains the classifier only in the experiments, and the results can be found in Table V. From these results, we find that: 1) When filtering the samples with high entropy only, the variant of training only classifier achieves higher performance than that training all the parameters. The test set inevitably contains biased samples, which may hinder the model's performance in the process of model adaptation. When performing the model adaptation, only updating the parameters of the classifier may be less affected by the biased samples than updating all the parameters, thus achieving higher performance. 2) When alleviating the bias issue in the adapting process, these two settings obtain higher performance, while training all the parameters performs better than that training the classifier only, which demonstrates the importance and necessity of alleviating the bias issue in the fully test-time adaptation setting. Moreover, these results further provide a guidance on the trade-off between accuracy and computational overhead.

5) Evaluation of Different Baseline Models: Our proposed method is model-agnostic. To demonstrate the effectiveness of our method on different baseline models, we perform experiments on the VQA-CP v1 and VQA-CP v2 datasets by using different baseline models (*i.e.,* UpDn [13], LXMERT [55], and ViLBERT [56]), and the experimental results are shown in Table IV. From these results, our TDS is able to promote the model performance regardless of the baseline models, which demonstrates that our TDS is model-agnostic, embodying the superiority of our method.

TABLE V
EFFECT OF THE TRAINABLE PARAMETERS ON THE MODEL PERFORMANCE IN TERMS OF ACCURACY (%)

Model	Components	VQA-CP v2 test set			
		All	Yes/No	Number	Other
UpDn [13]	-	40.94	43.87	12.90	47.09
TDS (CLS)	**ENT**	**41.41**	**43.90**	**15.06**	47.34
TDS	ENT	41.27	43.80	14.28	**47.36**
TDS (CLS)	ENT + BIAS	43.01	58.54	10.40	43.83
TDS	**ENT + BIAS**	**45.32**	**64.11**	**11.16**	**44.84**
TDS (CLS)	ALL	45.66	59.37	**10.56**	**48.11**
TDS	**ALL**	**46.33**	**62.55**	10.53	47.66

'CLS' denotes that we only train the parameters of the classifier, and fix the rest of the parameters. The baseline model is UpDn.

VI. CONCLUSION

In this paper, we have proposed a novel method named TDS to improve the VQA model performance at test time using only the test data without labels. Specifically, to alleviate the negative effect of the samples with high entropy, we set a threshold of entropy to filter the samples whose entropy is higher than the threshold. Moreover, when performing model adaptation during test time, the bias issue cannot be ignored. Thus, we first require to recognise the biased samples, and then remove the entropy of these samples. To find the biased samples, we consider constructing a negative sample for each sample, and regard the samples as biased samples if the output answers are the same when feeding the samples and the counterpart negative samples into the VQA model. Except for removing the entropy of biased samples, we also consider mitigating the bias in the VQA model by using biased samples, *i.e.,* minimising the possibility of predicting the answer of biased samples and their counterpart negative samples. Extensive experiments on the VQA-CP v2 dataset demonstrate the effectiveness of our TDS.

REFERENCES

[1] X. Zhang, F. Zhang, and C. Xu, "Explicit cross-modal representation learning for visual commonsense reasoning," *IEEE Trans. Multimedia*, vol. 24, pp. 2986–2997, 2021.

[2] Y. Zheng et al., "Modular graph attention network for complex visual relational reasoning," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 137–153.

[3] Y. Qi et al., "The road to know-where: An object-and-room informed sequential BERT for indoor vision-language navigation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 1655–1664.

[4] M. Tan, Z. Wen, L. Fang, and Q. Wu, "Transformer-based relational inference network for complex visual relational reasoning," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 1, pp. 1–23, 2023.

[5] G. Xu et al., "Towards accurate text-based image captioning with content diversity exploration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12637–12646.

[6] K. Nguyen, D. Dey, C. Brockett, and B. Dolan, "Vision-based navigation with language-based assistance via imitation learning with indirect intervention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12527–12537.

[7] L. Ke et al., "Tactical rewind: Self-correction via backtracking in vision-and-language navigation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6741–6749.

[8] R. Saqur and K. Narasimhan, "Multimodal graph networks for compositional generalization in visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 3070–3081.

[9] W. Guo et al., "Re-attention for visual question answering," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 91–98.

[10] R. Cadène, H. Ben-younes, M. Cord, and N. Thome, "MUREL: Multimodal relational reasoning for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1989–1998.

[11] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Don't just assume; look and answer: Overcoming priors for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4971–4980.

[12] S. Ramakrishnan, A. Agrawal, and S. Lee, "Overcoming language priors in visual question answering with adversarial regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1548–1558.

[13] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086.

[14] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–15.

[15] S. Niu et al. et al., "Efficient test-time model adaptation without forgetting," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 16888–16 905.

[16] M. Zhang, S. Levine, and C. Finn, "Memo: Test time robustness via adaptation and augmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 38629–38642.

[17] R. Cadène et al., "Rubi: Reducing unimodal biases for visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 839–850.

[18] S. Antol et al., "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2425–2433.

[19] F. Liu, J. Liu, Z. Fang, R. Hong, and H. Lu, "Visual question answering with dense inter-and intra-modality interactions," *IEEE Trans. Multimedia*, vol. 23, pp. 3518–3529, 2020.

[20] Y. Liu et al., "Depth-aware and semantic guided relational attention network for visual question answering," *IEEE Trans. Multimedia (TMM)*, to be published, doi: 10.1109/TMM.2022.3190686.

[21] B. Qin, H. Hu, and Y. Zhuang, "Deep residual weight-sharing attention network with low-rank attention for visual question answering," *IEEE Trans. Multimedia*, to be published, doi: 10.1109/TMM.2022.3173131.

[22] J. Yu et al., "Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval," *IEEE Trans. Multimedia (TMM)*, vol. 22, no. 12, pp. 3196–3209, Dec. 2020.

[23] T. Qian, J. Chen, S. Chen, B. Wu, and Y.-G. Jiang, "Scene graph refinement network for visual question answering," *IEEE Trans. Multimedia*, to be published, doi: 10.1109/TMM.2022.3169065.

[24] Y. Ding et al., "MuKEA: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5089–5098.

[25] J. Wu, J. Lu, A. Sabharwal, and R. Mottaghi, "Multi-modal answer validation for knowledge-based VQA," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2712–2721.

[26] M. Li and M. Moens, "Dynamic key-value memory enhanced multi-step graph reasoning for knowledge-based visual question answering," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 10983–10992.

[27] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6325–6334.

[28] H. Zhong et al., "Self-adaptive neural module transformer for visual question answering," *IEEE Trans. Multimedia*, vol. 23, pp. 1264–1273, 2020.

[29] L. Chen et al., "Counterfactual samples synthesizing for robust visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10797–10806.

[30] Y. Niu et al., "Counterfactual VQA: A cause-effect look at language bias," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12700–12710.

[31] N. Ouyang et al., "Suppressing biased samples for robust VQA," *IEEE Trans. Multimedia*, vol. 24, pp. 3405–3415, 2021.

[32] R. R. Selvaraju et al., "Taking a HINT: Leveraging explanations to make vision and language models more grounded," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2591–2600.

[33] J. Wu and R. J. Mooney, "Self-critical reasoning for robust visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8601–8611.

[34] X. Zhu et al., "Overcoming language priors with self-supervised learning for visual question answering," in *Proc. Int. Joint Conf. Artif. Intell.*, 2020, pp. 1083–1089.

[35] D. Teney et al., "On the value of out-of-distribution testing: An example of goodhart's law," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 407–417.

[36] Z. Wen, G. Xu, M. Tan, Q. Wu, and Q. Wu, "Debiased visual question answering from feature and sample perspectives," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 3784–3796.

[37] T. Gokhale, P. Banerjee, C. Baral, and Y. Yang, "MUTANT: A training paradigm for out-of-distribution generalization in visual question answering," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 878–892.

[38] Z. Wen, Y. Wang, M. Tan, Q. Wu, and Q. Wu, "Digging out discrimination information from generated samples for robust visual question answering," in *Proc. Findings Assoc. Comput. Linguistics: ACL*, 2023, pp. 1–17.

[39] Y. Sun et al., "Test-time training with self-supervision for generalization under distribution shifts," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9229–9248.

[40] Y. Liu et al., "TTT : When does self-supervised test-time training fail or thrive?" *Advances Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 21808–21820.

[41] A. Bartler, A. Bühler, F. Wiewel, M. Döbler, and B. Yang, "MT3: Meta test-time training for self-supervised test-time adaption," in *Proc. Int. Conf. Artif. Intell. Statist.*, vol. 151, 2022, pp. 3080–3090.

[42] M. Shu et al., "Test-time prompt tuning for zero-shot generalization in vision-language models," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 14274–14289.

[43] Q. Wang, O. Fink, L. V. Gool, and D. Dai, "Continual test-time domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7191–7201.

[44] M. Boudiaf, R. Müller, I. B. Ayed, and L. Bertinetto, "Parameter-free online test-time adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8334–8343.

[45] Z. Qiu et al., "Source-free domain adaptation via avatar prototype generation and adaptation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, pp. 2921–2927.

[46] S. Niu et al., "Towards stable test-time adaptation in dynamic wild world," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–27.

[47] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–16.

[48] J. Gao et al., "Back to the source: Diffusion-driven test-time adaptation," in *Proc. ICML Workshop Updatable Mach. Learn.*, 2022, pp. 1–19.

[49] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[50] N. Hansen et al., "Self-supervised policy adaptation during deployment," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–18.
[51] V. H. Pong, A. V. Nair, L. M. Smith, C. Huang, and S. Levine, "Offline meta-reinforcement learning with online self-supervision," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 17811–17829.
[52] Y. Li, M. Hao, Z. Di, N. B. Gundavarapu, and X. Wang, "Test-time personalization with a transformer for human pose estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 2583–2597.
[53] H. Liu et al., "Towards multi-domain single image dehazing via test-time training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5831–5840.
[54] Z. Chi, Y. Wang, Y. Yu, and J. Tang, "Test-time fast adaptation for dynamic scene deblurring via meta-auxiliary learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9137–9146.
[55] H. Tan et al., "LXMERT: Learning cross-modality encoder representations from transformers," in *Proc. Conf. Empirical Methods Natural Lang. Process.* K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., 2019, pp. 5099–5110.
[56] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 13–23.
[57] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[58] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.
[59] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

Zhiquan Wen received the B.E. degree in software engineering in 2019 from the School of Software Engineering, South China University of Technology, Guangzhou, China, where he is currently working toward the Ph.D. degree. His research interests include deep learning, and vision-and-language.

Shuaicheng Niu is currently working toward the Ph.D. degree with the South China University of Technology, Guangzhou, China. He has authored or coauthored papers in top venues, including ICML, ICLR, CVPR, ECCV, IJCAI, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON IMAGE PROCESSING, and IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING. His research interests are broadly in machine learning and mainly focus on domain adaptation, out-of distribution generalization, and automated machine learning. He is also a invited as a Reviewer for top-tier conferences and journals, including NeurIPS, ICML, CVPR, ICCV, ECCV, IEEE TRANSACTIONS ON IMAGE PROCESSING, and IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

Ge Li is currently a Professor with the School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, China. His research interests include image/video process and analysis, machine learning, digital communications, and signal processing.

Qingyao Wu received the B.S. degree in software engineering from the South China University of Technology, Guangzhou, China, in 2007, and the M.S. and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2009 and 2013, respectively. He is currently a Professor with the School of Software Engineering, South China University of Technology. His research interests include computer vision and data mining.

Mingkui Tan received the bachelor's degree in environmental science and engineering, the master's degree in control science and engineering from Hunan University, Changsha, China, in 2006 and 2009, respectively, and the Ph.D. degree in computer science from Nanyang Technological University, Singapore, in 2014.

From 2014 to 2016, he was a Senior Research Associate of computer vision with the School of Computer Science, University of Adelaide, Adelaide, SA, Australia. He is currently a Professor with the School of Software Engineering, South China University of Technology, Guangzhou, China. His research interests include machine learning, sparse analysis, deep learning and large-scale optimization.

Qi Wu received the M.Sc. and Ph.D. degrees in computer science from the University of Bath, Bath, U.K., in 2011 and 2015, respectively. His educational backgrounds primarily include computer science and mathematics. He is currently an Assistant Professor with The University of Adelaide, Adelaide, SA, Australia, where he is also an Associate Investigator with the Australia Centre for Robotic Vision (ACRV), Brisbane, QLD, Australia. His research interests include vision and language problems, image captioning, visual question answering, and visual dialog, etc.