# Overcoming Dual Drift for Continual Long-Tailed Visual Question Answering

## Supplementary Material

In this supplementary material, we will provide additional details on the aspects omitted in the main paper.

- **Section A. Additional Implementation details**: More construction details about the proposed Continual Long-Tailed Visual Question Answering (CLT-VQA).
- **Section B. Terminology Explanation**: Detailed explanation of the Neural Collapse (NC) and Optimal Transport (OT).
- **Section C. Theoretical Justification**: Detailed analysis of how balanced prototypes promote learning a balanced feature space.
- **Section D. Computation Efficiency**: Comprehensive computation efficiency comparison between our method and various continual learning approaches in terms of the computational costs (learnable parameters, training time, and memory consumption) and complexity analysis.
- **Section E. Compared Methods**: Detailed analysis about the compared algorithms, including the long-tailed learning methods (LDAM [5] and GCL [12]), and the continual learning approaches (EWC [11], MAS [2], (ER [6], DER [4], CVS [18], and VQACL [20]).
- **Section F. Additional Experimental Results** More experimental results to comprehensively validate the effectiveness of the proposed method, including fine-grained results across sub-tasks, hyperparameter selection for trade-off parameter $\lambda$ and memory size $M$, evaluations on a more challenging dataset (VQA-CP v2 [1]), and experiments on additional methods.

## A. Additional Implementation Details

CLT-VQA focuses on the continual learning paradigm and each sub-task is characterized by a long-tailed data distribution. To equip the model with diverse reasoning skills essential for real-world applications, we construct CLT-VQA by dividing two widely used VQA datasets, VQA v2 [8] and TDIUC [10], into distinct sub-tasks based on their question type annotations. Specifically, VQA v2 [8] is partitioned into nine sub-tasks, as illustrated in Tab. 1, encompassing *Recognition*, *Count*, *Color*, *Subcategory*, *Action*, *Commonsense*, *Type*, *Location*, and *Causal*. The *Judge* task is excluded to preserve label diversity, as nearly 95% of its samples belong to the "yes/no" class. Specifically, in the ordered scenario, the "Judge" task appears first due to its abundant training samples, which biases the model

Table 1. Task statistics of VQA v2 in the CLT-VQA setting.

| Task | Train | Test | Examples |
|---|---|---|---|
| Recognition | 144368 | 6185 | What is the train going over? What is the bird doing? |
| Count | 68131 | 2894 | How many giraffes are there? How many planes are in the picture? |
| Color | 56884 | 2435 | What color is his coat? What color is the bear? |
| Subcategory | 35624 | 1597 | What brand of beer is visible? What brand of soda is advertised? |
| Action | 35175 | 1448 | What is the person doing now? What is the man holding in his left hand? |
| Commonsense | 28107 | 1219 | Does the man look happy? Does the guy have a tattoo? |
| Type | 26388 | 1204 | What type of watercraft is that? What kind of room is this? |
| Location | 14840 | 696 | Where are the trucks? Where is the bird? |
| Causal | 6339 | 216 | Why is the man on the street? Why is the girl holding an umbrella? |

Table 2. Task statistics of TDIUC in the CLT-VQA setting.

| Task | Train | Test | Examples |
|---|---|---|---|
| Color | 133074 | 62490 | What color is in the bike? What color is the man's t-shirt? |
| Counting | 111857 | 52905 | How many boards are there? How many cups can be seen? |
| Object_R | 62862 | 30693 | What furniture is shown in the photo? What animal is shown in the picture? |
| Scene_R | 44674 | 22032 | What is the weather like? What season is the child dressed for? |
| Positional_R | 26042 | 12284 | What is to the left of cup? What is behind the calm water? |
| Sport_R | 21602 | 10042 | What sport is this? What sport is depicted in the picture? |
| Attribute | 19476 | 9200 | What is the sign made of? What is the ground made of? |
| Activity_R | 5848 | 2682 | What is the boy doing? What is the dog doing? |

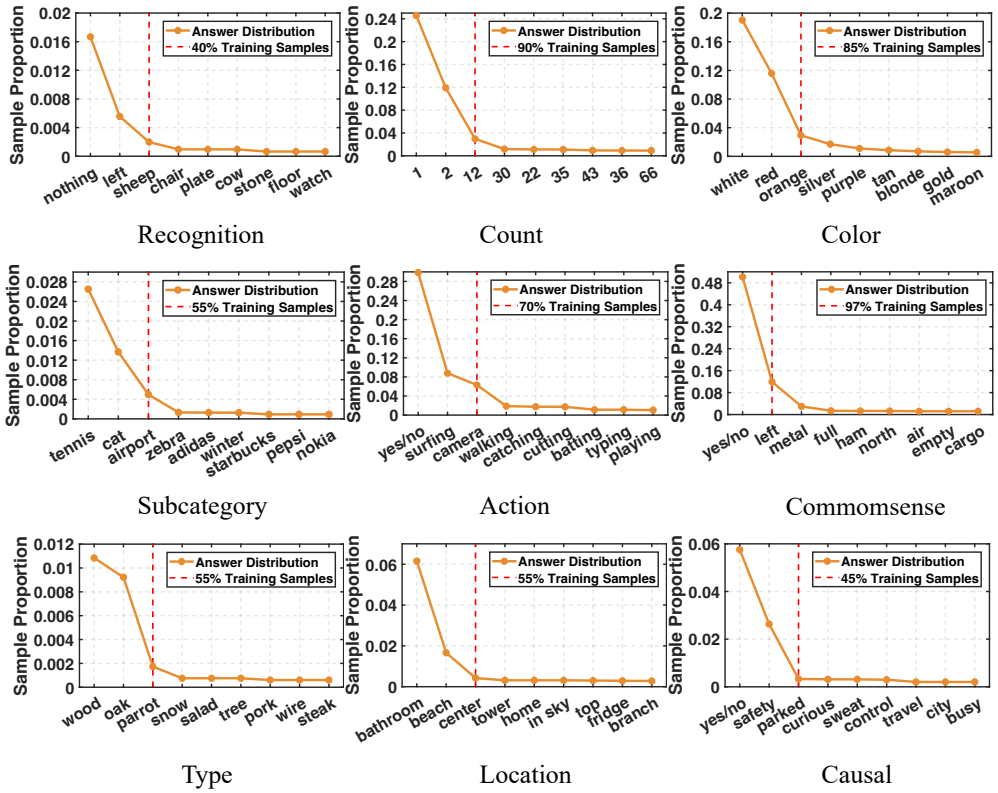



Figure 1. The long-tailed distribution in each sub-task of VQA v2.

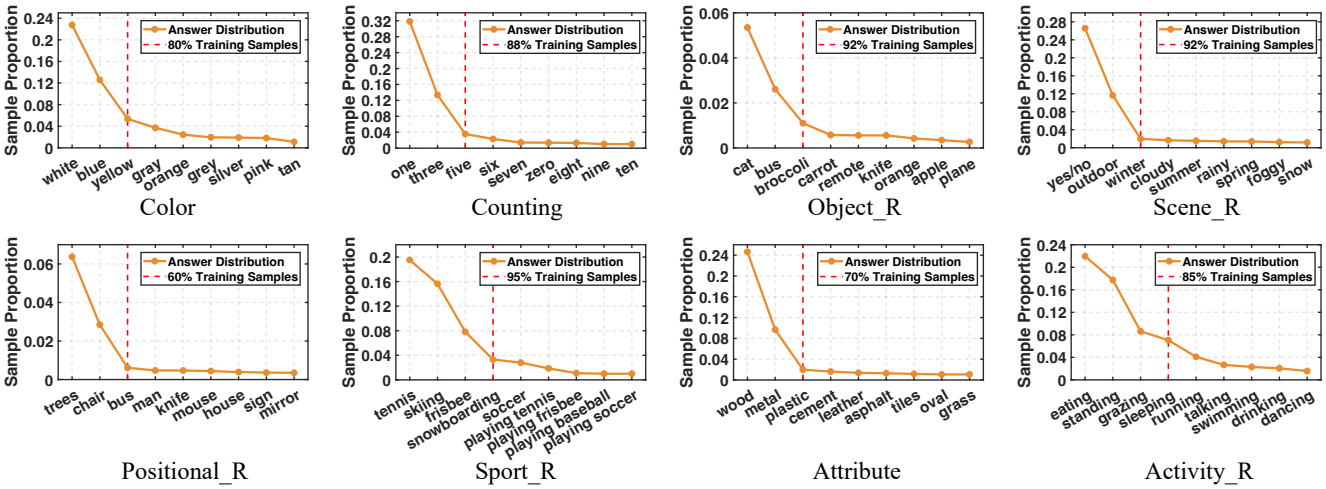



Figure 2. The long-tailed distribution in each sub-task of TDIUC.

toward binary classification and significantly impairs its ability to learn subsequent classes in later tasks. However, for comprehensive comparison, we also provide the experimental results on all ten tasks in this supplementary material (Section F.1). On TDIUC [10] dataset, we segment it into eight sub-tasks, as detailed in Tab. 2, which include *Color*, *Counting*, *Object_Recognition (Object_R)*, *Scene_Recognition (Scene_R)*, *Positional_Reasoning (Positional_R)*, *Sport_Recognition (Sport_R)*, *Attribute*, and *Activity_Recognition (Activity_R)*.

Each sub-task in both datasets exhibits a long-tailed data distribution, as depicted in Fig. 1 and Fig. 2. In both figures, the red dashed line indicates a dynamically set threshold determined by the number of samples per class. Classes below this threshold are identified as minority classes. For clarity and to reduce visual clutter, only nine

representative answers are displayed in the distributions.

## B. Terminology Explanation

**Neural collapse (NC)**: Papyan *et al*. [13] revealed the neural collapse phenomenon, where last-layer features and classifier vectors converge to form a simple Equiangular Tight Frame (ETF) at the terminal phase of training (after 0 training error rate) on balanced datasets. A standard simplex ETF is a collection of vectors in $\mathbb{R}^K$ that satisfy the following properties:

$$W^{'} = \sqrt{\frac{K}{K-1}}\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^{\top}\right), \tag{1}$$

where $W^{'} = \left[w_1^{'}, \ldots, w_K^{'}\right] \in \mathbb{R}^{K\times K}$ is a matrix composed of $K$ vectors, $\mathbf{I}_K$ denotes the identity matrix, and $\mathbf{1}_K$ is an all-ones vector. In this manner, $W'$ itself, or transformations obtained by applying an orthogonal transformation from the left (e.g., $\hat{W} = UW'$, where $\mathbf{U} \in \mathbb{R}^{d\times K}$ $(d \geq K)$ is an orthogonal rotation matrix satisfying $\mathbf{U}^{\top}\mathbf{U} = \mathbf{I}_K$) could preserve the same length and angle for any pair of vectors. In other words, all vectors in $W^{'}$ have an equal $\ell_2$ norm and the same pair-wise angle as follows:

$$w_i^{'\top} w_j^{'} = \frac{K}{K-1}\delta_{i,j} - \frac{1}{K-1}, \forall i, j \in \{1, \ldots, K\}, \tag{2}$$

where $\delta_{i,j}$ equals to 1 when $i = j$ and 0 otherwise. The pairwise angle $-\frac{1}{K-1}$ is the maximal equiangular separation of $K$ vectors in $\mathbb{R}^d$.

As pointed by reference [19], the NC phenomenon includes four important geometric properties as follows:

1. The last-layer features of the same class collapse to their class mean: $\Sigma_W \to 0$, and $\Sigma_W = \mathrm{Avg}_{i,k}\left\{(z_{k,i} - z_k)(z_{k,i} - z_k)^T\right\}$, where $z_{k,i}$ is the feature of sample $i$ in class $k$, and $z_k$ is the class mean of representations from the class $k$.
2. The class means ($\hat{z}_k$, $1 \leq k \leq K$) centered at their global mean converge to the vertices of a simplex ETF, and $\hat{z}_k = (z_k - z_G) / \|z_k - z_G\|$ with $z_G$ is the global mean of the last-layer features for all samples.
3. The classifier prototypes could also converge to the same simplex ETF like the class means $\hat{z}_k$, *i.e*., $\hat{w}_k^{'} = w_k^{'} / \left\|w_k^{'}\right\| = \hat{z}_k, 1 \leq k \leq K$, where $w_k^{'}$ is the classifier prototype of the $k$-th class.
4. The learned classifier behaves like the nearest classifier: $\arg\max_k \left\langle z, w_k^{'}\right\rangle = \arg\min_k \|z - z_k\|$, where $\langle\cdot\rangle$ is the inner product operator, $z$ is the last-layer feature of a sample for prediction.

Therefore, by initializing our VQA classifier using $\hat{W}$, the classifier inherently retains the aforementioned properties, which contribute to improved performance in addressing the challenges posed by long-tailed distributions. However, as discussed in the main paper, we identify two limitations: first, initializing a simplex ETF classifier with fixed directions may hinder the model's generalization capability. Second, a simplex ETF can only exist when the feature dimension $d$ exceeds the number of classes $K$, which is often impractical in real-world VQA scenarios where the number of classes varies significantly. Therefore, we introduce the learnable orthogonal matrix $U^*$ to further refine $\hat{W}$, *i.e*., $W^* = U^*W^{'}$, which not only ensures that $W^*$ preserves the equiangular property of $W^{'}$ but also keeps it optimal for each class by dynamically learning the appropriate prototype directions of $W^{'}$, thereby enhancing the model's generalization and improving its discriminative ability under the long-tailed distribution.

**Optimal Transport (OT)**: OT [14] typically aims to find the most cost-effective way to transform one distribution into another, which is achieved by calculating a transportation plan that minimizes the total transportation cost. This minimized cost is known as the OT distance. From the matching perspective, OT provides a geometrically meaningful distance between probability distributions, showcasing its effectiveness in multiple domains of machine learning [9, 17]. Suppose we have two discrete distributions in the same arbitrary space: $\alpha = \sum_{i=1}^{n} a_i\delta_{x_i}$ and $\beta = \sum_{j=1}^{m} b_j\delta_{y_j}$, where $\delta$ is a Dirac function, and $a_i$ and $b_j$ are the weight. The discrete optimal transport problem can be formulated as follows:

$$T^* = \underset{\mathbf{T}\in\Pi(\alpha,\beta)}{\arg\min} \sum_{i=1}^{n}\sum_{j=1}^{m} T_{ij}C_{ij}, , \tag{3}$$

where $T^*$ is the optimal transport plan learned to minimize the total distance between two probability vectors. The transport probability matrix $T \in \mathbb{R}_+^{n\times m}$, which satisfies $\Pi(\alpha,\beta) := \left\{T \mid \sum_{i=1}^{n} T_{ij} = b_j, \sum_{j=1}^{m} T_{ij} = a_i\right\}$. The cost matrix $C \in \mathbb{R}_+^{n\times m}$ represents the distances between representation $x_i$ and $y_j$ .

In our paper, we minimize the OT distance between the multimodal feature distribution over each sample $F$ from the backbone network and the corresponding balanced classifier prototype distribution $W^*$ in the BCP module. This alignment of multimodal features with their respective classifier prototypes helps in mitigating the catastrophic forgetting caused by feature drift in CLT-VQA.

## C. Theoretical Justification

**Theorem 1.** Given a $k$-equiangular ETF $\left\{w_i'\right\}_{i=1}^{K}$, where the inner product $\left\langle w_i', w_j'\right\rangle = k$ for any $i, j$ $(i \neq j)$. $\bar{w}' = \frac{1}{K}\sum_{i=1}^{K} w_i'$ is the mean of $\left\{w_i'\right\}_{i=1}^{K}$. Consequently, $\left\{w_i' - \bar{w}'\right\}_{i=1}^{K}$ forms a regular simplex, and possess the following properties:

(1) *zero mean*: $\sum_{i=1}^{K}\left(w_i' - \bar{w}'\right) = 0$;

(2) *equalnorm*: $\forall i,\ \left\|w_i' - \bar{w}'\right\| = \sqrt{\mathcal{M}(K,k)}$;

(3) *equiangular*: $\forall i \neq j,\ \left\langle w_i' - \bar{w}', w_j' - \bar{w}' \right\rangle = \mathcal{N}(K,k)$.

In the following, we provide a detailed proof of the three properties.

***Proof of property (1):*** The *zero mean* property can be proofed as follows:

$$\sum_{i=1}^{K}\left(w_i' - \bar{w}'\right) = \sum_{i=1}^{K} w_i' - K\bar{w}' = 0. \tag{4}$$

***Proof of property (2):*** The *equalnorm* property can be proofed as follows:

$$\begin{aligned}
\left\|w_i' - \bar{w}'\right\|^2 &= \left\langle w_i', w_i' \right\rangle - 2\left\langle w_i', \bar{w}' \right\rangle + \left\langle \bar{w}', \bar{w}' \right\rangle \\
&= 1 - \frac{2}{K}\sum_{k=1}^{K}\left\langle w_i', w_k' \right\rangle + \frac{1}{K^2}\sum_{k=1}^{K}\sum_{n=1}^{K}\left\langle w_k', w_n' \right\rangle \\
&= 1 - 2\frac{1+(K-1)k}{K} + \frac{K(1+(K-1)k)}{K^2} \\
&= 1 - \frac{1+(K-1)k}{K}.
\end{aligned} \tag{5}$$

***Proof of property (3):*** Given any $i, j (i \neq j)$, we have:

$$\begin{aligned}
\left\langle w_i' - \bar{w}', w_j' - \bar{w}' \right\rangle &= \left\langle w_i', w_j' \right\rangle - \left\langle \bar{w}', w_j' \right\rangle \\
&\quad -\left\langle w_i', \bar{w}' \right\rangle + \left\langle \bar{w}', \bar{w}' \right\rangle \\
&= k - 2\frac{1+(K-1)k}{K} + \frac{K(1+(K-1)k)}{K^2} \\
&= k - \frac{1+(K-1)k}{K}.
\end{aligned} \tag{6}$$

For simplicity, we denote $1 - \frac{1+(K-1)k}{K}$ as $\mathcal{M}(K,k)$ and $k - \frac{1+(K-1)k}{K}$ as $\mathcal{N}(K,k)$. These will be used in the following derivations. Next, the focus is on proving that within the BCP module, the balanced prototypes encourage the multi-modal representation $f$ to gradually converge toward its corresponding classification prototypes under the ETF structure. Recall the optimization objective in Eq. (4) in our main paper:

$$\begin{aligned}
\min_{\theta,\psi} \mathcal{L}^*(v,q;\theta,\psi) &= \frac{1}{|D^{(t)}| + |M|} \sum_{(v,q,y)\in D^{(t)}\cup M} \ell\left(y^*(v,q), y\right) \\
&= \frac{1}{N^{(t)}} \sum_{k=1}^{K^{(t)}} \sum_{i=1}^{n_k} \left[ -\log \frac{\exp\left(\left[\text{logit}(v_{k,i}^{(t)}, q_{k,i}^{(t)})\right]_{y_i}\right)}{\sum_{y=1}^{K}\exp\left(\left[\text{logit}(v_{k,i}^{(t)}, q_{k,i}^{(t)})\right]_{y}\right)} \right],
\end{aligned} \tag{7}$$

where $(v_{k,i}^{(t)}, q_{k,i}^{(t)})$ denotes the $i$-th sample from class $k$ in task $t$, with $1 \leq t \leq T$, $1 \leq k \leq K^{(t)}$, and $1 \leq i \leq n_k$. Here, $n_k$ represents the number of samples in class $k$, $K^{(t)}$ is the total number of classes in task $t$, and $N^{(t)}$ is the total number of samples in task $t$, *i.e.*, $N^{(t)} = \sum_{k=1}^{K^{(t)}} n_k$. Since the problem is separable across $T$ tasks, we only analyze the $t$-th task and omit the superscript $(t)$ for simplicity. The objective in Eq. (7) can be simplified as follows:

$$\min_{\theta,\psi} \mathcal{L}^*(v,q;\theta,\psi) = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\left[\text{logit}((v_i,q_i))\right]_{y_i}\right)}{\sum_{y=1}^{K}\exp\left(\left[\text{logit}((v_i,q_i))\right]_{y}\right)}, \tag{8}$$

where $\text{logit}((v_i,q_i)) = W^* f_i = W^{'} U^* f_i$, and $\theta$ and $\psi$ denote the parameters of the feature extractor and $U^*$, respectively. $U^*$ is constrained to be an orthogonal matrix. $W' = \left[w_1'^T, \cdots, w_K'^T\right]^T$ is a pre-assigned ETF that remains balanced throughout continual learning, with each row vector $w_i^{'}$ originating from a $k$-equiangular ETF. Then, Eq. (8) can be reformulated as follows:

$$\begin{aligned}
&-\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\left[\text{logit}((v_i,q_i))\right]_{y_i}\right)}{\sum_{y=1}^{K}\exp\left(\left[\text{logit}((v_i,q_i))\right]_{y}\right)} \\
&= \frac{1}{N}\sum_{i=1}^{N} \log\Bigg(1 + \sum_{\substack{j\neq y_i \\ j\in[K]}} \exp\Big(\left[\text{logit}((v_i,q_i))\right]_j \\
&\qquad -\left[\text{logit}((v_i,q_i))\right]_{y_i}\Big)\Bigg) \\
&\overset{\text{C1}}{\geq} \frac{1}{N}\sum_{i=1}^{N} \log\left(1 + (K-1)\exp\left[\frac{1}{K-1}\sum_{\substack{j\neq y_i \\ j\in[K]}} \Delta_j^{(i)}\right]\right),
\end{aligned} \tag{9}$$

where $\Delta_j^{(i)} = \left[\text{logit}\left((v_i,q_i)\right)\right]_j - \left[\text{logit}\left((v_i,q_i)\right)\right]_{y_i}$. The inequality C1 follows from Jensen's inequality, and the equality holds when $\forall i \in [N], j \in [K]\,(j \neq y_i), \exists R_i \in \mathbb{R}, \Delta_j^{(i)} = R_i$. Due to the convexity of the function $x \to \log(1+\exp(x))$, Eq. (9) still follows from Jensen's inequality and can be reformulated as follows:

$$\begin{aligned}
&\overset{\text{C2}}{\geq} \log\left(1 + (K-1)\exp\left[\frac{1}{(K-1)N}\sum_{i=1}^{N}\sum_{\substack{j\neq y_i \\ j\in[K]}} \Delta_j^{(i)}\right]\right) \\
&= \log\left(1 + (K-1)\exp\left[\frac{1}{(K-1)N}\sum_{i=1}^{N}\sum_{j\in[K]} \Delta_j^{(i)}\right]\right).
\end{aligned} \tag{10}$$

The inequality C2 holds when $\forall i \in [N], \exists R \in \mathbb{R}, \sum_{j\neq y_i, j\in[K]} \Delta_j^{(i)} = R$. We first showcase the role of the balanced vectors in the ETF structure, $W^{'}$, and can rewrite the expression as follows:

$$\begin{aligned}
\text{logit}((v_i,q_i)) &= W^{'} f_i = W' U^* f_i \\
&= \left[w_1' U^* f_i, \ldots, w_K' U^* f_i\right]^T \\
&= \left[\left\langle w_1'^T, U^* f_i \right\rangle, \ldots, \left\langle w_K'^T, U^* f_i \right\rangle\right]^T.
\end{aligned} \tag{11}$$

Due to the fact that the function $x \to \log(1+\exp(x))$

is monotonically increasing, we now attempt to bound $\sum_{i=1}^{N}\sum_{j\in[K]}\Delta_j^{(i)}$ based on Eq. (11) and can be reformulated as follows:

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{j\in[K]}\Delta_j^{(i)} \\
&=\sum_{i=1}^{N}\sum_{j\in[K]}\Big([\text{logit}((v_i,q_i))]_j - [\text{logit}((v_i,q_i))]_{y_i}\Big) \\
&=\sum_{i=1}^{N}\sum_{j\in[K]}\Big(\langle w_j'^T, U^*f_i\rangle - \langle w_{y_i}'^T, U^*f_i\rangle\Big) \\
&=K\sum_{i=1}^{N}\langle(\bar{W}'-w_{y_i}'^T), U^*f_i\rangle \\
&\overset{\text{C3}}{\geq} -K\sum_{i=1}^{N}\|(\bar{W}'-w_{y_i}'^T)\|\|U^*f_i\|.
\end{aligned} \tag{12}$$

The inequality C3 follows from the Cauchy-Schwarz inequality, and the equality holds if and only if when $\forall i \in [N],\ \exists\lambda_i\in\mathbb{R}^+,\ U^*f_i=\lambda_i(w_{y_i}'^T-\bar{W}')$. Assuming that the multimodal feature $f$ has a maximum norm of $\rho$, we can further derive the following inequality based on Eq. (12):

$$\overset{\text{C4}}{\geq} -K\rho\sum_{i=1}^{N}\|(\bar{W}'-w_{y_i}'^T)\|. \tag{13}$$

The inequality in Eq. (13) holds under the assumption of the feature extractor, and the equality holds if and only if $\forall i \in [N], \|U^*f_i\| = \rho$. We then derive the conditions C1–C4 under which equality holds when the objective function converges to its minimum value, providing insights into the characteristics of the learned multimodal features $f$. Given a VQA sample $(v_i, q_i, y_i)$, starting from C3–C4, we have the following:

$$\|U^*f_i\| = \lambda_i\|w_{y_i}'^T-\bar{W}'\| = \rho, \tag{14}$$

and for any $i \in [N]$, we have the following:

$$\lambda_i = \frac{\rho}{\|w_{y_i}'^T-\bar{W}'\|} = \frac{\rho}{\sqrt{\mathcal{M}(K,k)}}. \tag{15}$$

According to Theorem 1, we know that the sequence of vector $\langle w_y'^T-\bar{W}'\rangle_{y=1}^K$ is equalnorm and equiangular. Therefore, the learned multimodal features $f$ will ultimately also be equalnorm and equiangular. Then return to the logit of sample $(v_i, q_i, y_i)$, we have:

$$\begin{aligned}
\text{if } y\neq y_i,\ \big[\text{logit}((v_i,q_i))\big]_y &= \langle w_y'^T, \lambda_i(w_{y_i}'^T-\bar{W}')\rangle \\
&= \lambda_i\mathcal{N}(K,k) = \rho\frac{\mathcal{N}(K,k)}{\sqrt{\mathcal{M}(K,k)}}, \\
\text{if } y=y_i,\ \big[\text{logit}((v_i,q_i))\big]_y &= \langle w_{y_i}'^T, \lambda_i(w_{y_i}'^T-\bar{W}')\rangle \\
&= \lambda_i\mathcal{M}(K,k) = \rho\sqrt{\mathcal{M}(K,k)}.
\end{aligned} \tag{16}$$

According to Eq. (16), we derive the values of $R_i$ in C1 and $R$ in C2:

$$\begin{aligned}
R_i &= \rho\frac{\mathcal{N}(K,k)}{\sqrt{\mathcal{M}(K,k)}} - \rho\sqrt{\mathcal{M}(K,k)}, \\
R &= \rho(K-1)\Big(\frac{\mathcal{N}(K,k)}{\sqrt{\mathcal{M}(K,k)}} - \sqrt{\mathcal{M}(K,k)}\Big).
\end{aligned} \tag{17}$$

## D. Computation Efficiency

In the BCP module, the orthogonal projection matrix $U^*$ is orthogonalized only once at the beginning of the training process. During the training phase, it is updated in the same way as the baseline classifier, following the standard training procedure. Moreover, the ETF structure, as defined in Sec. B, is initialized before training based on the specific number of classes and serves as a fixed target that remains unchanged throughout the training process. Thus, the main computational cost comes from the OT distance in MFA. Next, we will analyze our method in terms of computational cost and complexity, and the results are shown in Tab. 3.

**Computational cost**: Tab. 3 presents a comparative analysis of computational costs associated with our method and other continual learning approaches in terms of the number of learned parameters, training time, and memory consumption. The results indicate that our method achieves the lowest parameter count among the evaluated approaches while introducing only a marginal increase in training time and memory usage. Consequently, we believe our approach maintains a reasonable computational cost, particularly considering its performance advantages.

**Complexity**: To approximate the OT distance between two discrete distributions of size $n$, the time complexity bound scales as $O\left(n^2\log(n)/\epsilon^2\right)$ to achieve $\epsilon$-accuracy with Sinkhorn's algorithm, as demonstrated by [3, 7]. In this paper, for each sample $(v_i, q_i)$, we push its multimodal feature distribution toward the corresponding balanced classifier prototype distribution by minimizing their OT distance. Thus, the sample-wise time complexity bound scales as $O\left(K^2\log(K)/\epsilon^2\right)$, where $K$ is the number of classifier prototypes. The number of prototypes $K$ is dataset-specific; for the VQA v2 dataset, $K$=3129, and for the TDIUC dataset, $K$=1589, both of which are much smaller than the number of features.

Table 3. Computational cost of our method and compared methods. Param: Learnable parameters. Time: Training time. Memory: Memory usage.

| Methods | VQA v2 | | | TDIUC | | |
|---|---|---|---|---|---|---|
| | Param | Time | Memory | Param | Time | Memory |
| EWC [11] | 230.54M | 2.79h | **12.00G** | 228.17M | 2.02h | **10.20G** |
| MAS [2] | 230.54M | **2.26h** | 14.10G | 228.17M | **1.93h** | 13.40G |
| DER [4] | 230.54M | 4.07h | 16.10G | 228.17M | 3.27h | 14.10G |
| ER [6] | 230.54M | 3.60h | 12.20G | 228.17M | 2.32h | 10.60G |
| CVS [18] | 231.71M | 4.36h | 15.40G | 229.35M | 3.50h | 13.00G |
| VQACL [20] | 231.71M | 3.80h | 13.60G | 229.35M | 2.41h | 11.20G |
| Ours | **226.94M** | 3.90h | 14.30G | **225.76M** | 2.48h | 11.80G |

## E. Compared Methods

To evaluate the effectiveness of our approach, we conduct comparisons against several state-of-the-art methods, spanning both long-tailed learning and continual learning paradigms. The long-tailed learning methods include LDAM [5] and GCL [12], and the continual learning approaches comprise regularization-based methods (EWC [11], MAS [2]), replay-based methods (ER [6], DER [4], CVS [18]), and the VQA-specific model VQACL [20]. To ensure a fair comparison, all methods are implemented using their official code repositories and integrated with the encoder-decoder architecture introduced in Sec. 3.2 in our main paper. Specifically,

**LDAM** [5] introduces a modified loss function to enhance model generalization on long-tailed datasets by encouraging larger margins for minority classes, effectively addressing class imbalance. By adjusting the soft margin loss based on label distribution, LDAM [5] applies stronger regularization to minority classes while maintaining competitive performance on majority classes, ensuring a balanced trade-off across the dataset.

**GCL** [12] addresses the challenges of distorted embedding spaces and biased classifiers in long-tailed datasets. By introducing larger Gaussian noise to the logits of tail classes, GCL effectively pushes tail class samples further from decision boundaries. It mitigates embedding space distortion and improves classifier calibration, thereby enhancing the representation and recognition of tail class samples.

**EWC** [11] is a regularization method and remembers old tasks by selectively slowing down learning on the parameters that are important for these tasks. To achieve it, EWC uses the Fisher Information Matrix [15] to estimate the importance of each parameter and adds an auxiliary $L_2$ loss between the important parameters learned from the new task and old tasks.

**MAS** [2] is also a regularization method and discourages big changes in parameters that are important for previous tasks through an additional $L_2$ loss. To estimate the importance of a parameter, MAS measures how sensitive the predicted output function is to a change in this parameter.

**ER** [6] is a replay-based approach and randomly stores visited examples in a fix-sized memory. At each training step, it randomly samples these stored examples for retraining. Consistent with our method, the memory size of ER is set to 5,000 for VQA v2 [8] and 5,000 for TDIUC [10]. Since ER is well-established and simple to implement, we utilize it as the baseline of our proposed approach.

**DER** [4] belongs to replay-based methods and adopts reservoir sampling [16] to decide examples to store and replace from the replayed memory. Specifically, the reservoir algorithm ensures each visited example has the same probability to be stored in the memory. Based on the memory, DER designs a dark experience-based knowledge distillation strategy to match the network's output logits sampled throughout the training process, which encourages the network to mimic its original responses for past examples. In our experiments, the memory size is set to 5,000 for VQA v2 [8] and TDIUC [10].

**CVS** [18] is a replay-based method and considers the feature compatibility between the ongoing and previous data. To model the feature consistency and mitigate forgetting, it designs a neighbor-session model coherence loss and an inter-session data coherence loss. We suggest readers check Wan *et al*. [18] for more details about these two losses. As in our method, the memory size of CVS is set to 5,000 for VQA v2 [8] and TDIUC [10].

**VQACL** [20] is a replay-based representation learning method tailored for continual VQA. While VQACL is specifically designed for continual VQA, the challenges and objectives in our work are fundamentally different. Beyond continual learning in VQA, our study primarily tackles the long-tailed distribution problem in VQA systems, which remains an understudied yet crucial aspect. Furthermore, VQACL introduces a prototype learning module that leverages two complementary feature types: sample-specific (SS) features, which encapsulate task-specific information, and sample-invariant (SI) features, which capture stable and generalizable knowledge. By integrating these features, VQACL enables the model to effectively mitigate catastrophic forgetting while adapting to new tasks. To address catastrophic forgetting while tackling the long-tailed distribution, we propose two modules. First, the Balanced Classifier Prototype (BCP) Learning module addresses inner-task prototype drift caused by the long-tailed data distribution, ensuring balanced class representation across different categories. Second, the Multi-modal Feature Alignment (MFA) module mitigates inter-task feature drift by minimizing deviations between updated feature representations and their corresponding classifier prototypes, thus combating catastrophic forgetting. Similar to our approach, the memory size is set to 5,000 for VQA v2 [8] and TDIUC [10].

Table 4. The VQA performance (%) on the standard test set of VQA v2 under the CLT-VQA setting across 9 tasks in the ordered scenario. The memory size in the replay-based methods is 5,000. The best results are highlighted in bold.

| Method | Recognition | Count | Color | Subcategory | Action | Commonsense | Type | Location | Causal | AP |
|---|---|---|---|---|---|---|---|---|---|---|
| Joint | 35.55 | 39.51 | 61.8 | 52.94 | 62.23 | 71.49 | 42.03 | 34.56 | 16.55 | 46.30 |
| Vanilla | 2.22 | 0.26 | 0.04 | 24.15 | 31.66 | 62.65 | 1.36 | 6.66 | 13.50 | 15.83 |
| LDAM [5] | 2.87 | 0.24 | 0.03 | 26.79 | 34.47 | 64.41 | 0.95 | 4.47 | 13.70 | 16.44 |
| GCL [12] | 2.92 | 0.27 | 0.05 | 27.97 | 31.43 | 63.44 | 2.25 | 7.99 | 12.50 | 16.54 |
| EWC [11] | 18.10 | 0.27 | 21.21 | 34.56 | 43.82 | 63.28 | 12.62 | 11.10 | 12.15 | 24.12 |
| MAS [2] | 14.50 | 0.26 | 25.95 | 35.99 | 44.94 | 65.32 | 11.62 | 9.05 | 11.85 | 24.39 |
| DER [4] | 11.58 | 30.75 | 25.01 | 38.00 | 45.51 | 67.08 | 17.84 | 14.86 | 14.75 | 29.49 |
| ER [6] | 18.30 | 29.12 | 38.50 | 43.57 | 53.59 | 65.11 | 23.02 | 22.96 | 14.55 | 34.30 |
| CVS [18] | 15.18 | 30.65 | 28.19 | 41.56 | 47.82 | 58.29 | 19.75 | 20.65 | 13.91 | 30.67 |
| VQACL [20] | 22.60 | 30.02 | 50.12 | 47.14 | 53.62 | 66.27 | 30.82 | 27.81 | 13.45 | 37.98 |
| Ours | **25.71** | **32.76** | **52.73** | **50.46** | **55.03** | **68.73** | **34.92** | **31.16** | **14.75** | **40.69** |

Table 5. The VQA performance (%) on the standard test set of VQA v2 under the CLT-VQA setting across 9 tasks in the random scenario. The memory size in the replay-based methods is 5,000. The best results are highlighted in bold.

| Method | Recognition | Location | Commonsense | Count | Action | Color | Type | Subcategory | Causal | AP |
|---|---|---|---|---|---|---|---|---|---|---|
| Joint | 35.55 | 34.56 | 71.49 | 39.51 | 62.23 | 61.80 | 42.03 | 52.94 | 16.55 | 46.30 |
| Vanilla | 3.98 | 5.04 | 59.20 | 0.20 | 18.41 | 0.04 | 3.80 | 30.13 | 15.80 | 15.18 |
| LDAM [5] | 1.73 | 1.91 | 66.62 | 0.27 | 30.42 | 0.04 | 1.61 | 29.96 | 12.70 | 16.14 |
| GCL [12] | 2.06 | 0.83 | 66.13 | 0.25 | 35.23 | 0 | 1.85 | 29.62 | 12.15 | 16.46 |
| EWC [11] | 17.52 | 8.23 | 55.19 | 0.33 | 37.00 | 22.63 | 15.55 | 32.61 | 11.15 | 22.25 |
| MAS [2] | 16.83 | 9.25 | 57.04 | 2.11 | 36.73 | 24.11 | 16.32 | 38.04 | 9.45 | 23.32 |
| DER [4] | 11.47 | 10.87 | 66.17 | 31.49 | 44.57 | 31.88 | 14.03 | 41.46 | 14.55 | 29.61 |
| ER [6] | 18.67 | 25.60 | 64.54 | 28.10 | 52.25 | 32.77 | 22.46 | 43.86 | 12.05 | 33.37 |
| CVS [18] | 15.13 | 25.29 | 59.18 | 31.04 | 50.12 | 30.55 | 18.82 | 40.45 | 12.65 | 31.47 |
| VQACL [20] | 21.92 | 28.72 | 63.32 | 31.63 | 53.15 | 50.50 | 31.84 | 46.71 | 13.20 | 37.89 |
| Ours | **25.59** | **33.39** | **67.00** | **32.28** | **55.18** | **52.79** | **34.42** | **49.13** | **17.95** | **40.86** |

Table 6. The VQA performance (%) on the standard test set of VQA v2 under the CLT-VQA setting across 10 tasks in the random scenario. The memory size in the replay-based methods is 5,000. The best results are highlighted in bold.

| Method | Recognition | Location | Judge | Commonsense | Count | Action | Color | Type | Subcategory | Causal | AP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Joint | 34.75 | 33.76 | 67.48 | 68.69 | 38.71 | 61.43 | 61.00 | 41.23 | 52.14 | 15.75 | 47.49 |
| Vanilla | 1.55 | 3.29 | 55.23 | 65.9 | 0.27 | 32.04 | 0.05 | 0.29 | 25.78 | 12.75 | 19.72 |
| LDAM [5] | 1.72 | 2.36 | 57.32 | 66.1 | 0.34 | 33.2 | 0.04 | 1.32 | 27.33 | 13.7 | 20.34 |
| GCL [12] | 2.32 | 1.73 | 59.32 | 66.3 | 0.33 | 35.3 | 0.07 | 1.98 | 29.34 | 13.79 | 21.05 |
| EWC [11] | 16.09 | 6.55 | 52.68 | 60.03 | 0.12 | 41.99 | 20.64 | 12.61 | 31.61 | 12.20 | 25.45 |
| MAS [2] | 14.02 | 10.10 | 56.97 | 63.53 | 0.18 | 44.40 | 17.08 | 14.60 | 37.18 | 12.70 | 27.08 |
| DER [4] | 12.06 | 10.00 | 58.49 | 67.47 | 28.54 | 45.76 | 38.79 | 15.69 | 42.43 | 14.3 | 33.35 |
| ER [6] | 15.71 | 24.32 | 60.62 | 66.74 | 29.64 | 49.7 | 32.32 | 19.11 | 41.86 | 12.5 | 35.25 |
| CVS [18] | 15.15 | 23.70 | 60.13 | 61.93 | 30.87 | 49.20 | 28.76 | 17.23 | 39.60 | 12.65 | 33.92 |
| VQACL [20] | 19.81 | 25.83 | 62.47 | 65.35 | 31.78 | 53.02 | 41.55 | 24.33 | 45.72 | 14.5 | 38.44 |
| Ours | **24.98** | **28.25** | **63.48** | **68.81** | **32.19** | **55.88** | **47.64** | **31.36** | **46.79** | **14.75** | **41.41** |

Table 7. The VQA performance (%) on the standard test set of TDIUC under the CLT-VQA setting in the ordered scenario. The memory size in the replay-based methods is 5,000. The best results are highlighted in bold.

| Method | Color | Counting | Object_R | Scene_R | Positional_R | Sport_R | Attribute | Activity_R | AP |
|---|---|---|---|---|---|---|---|---|---|
| Joint | 58.67 | 50.21 | 87.08 | 92.21 | 29.84 | 95.49 | 46.97 | 49.79 | 63.78 |
| Vanilla | 0 | 0 | 0 | 0 | 0.04 | 0 | 0 | 31.05 | 3.89 |
| LDAM [5] | 0 | 0 | 0 | 0 | 0.04 | 0 | 0 | 33.93 | 4.25 |
| GCL [12] | 0 | 0 | 0 | 0 | 0.04 | 0 | 0 | 34.04 | 4.26 |
| EWC [11] | 13.68 | 22.86 | 24.17 | 42.70 | 0.34 | 21.27 | 6.97 | 33.93 | 20.74 |
| MAS [2] | 27.20 | 31.72 | 26.46 | 70.04 | 2.43 | 27.23 | 2.40 | 18.21 | 25.71 |
| DER [4] | 38.25 | 39.10 | 55.18 | 85.88 | 8.77 | 59.31 | 17.03 | 26.39 | 41.24 |
| ER [6] | 40.85 | 38.88 | 71.25 | 87.02 | 12.52 | 63.51 | 36.09 | 38.98 | 48.64 |
| CVS [18] | 41.00 | 34.41 | 54.43 | 81.85 | 11.03 | 73.46 | 33.23 | 33.16 | 45.32 |
| VQACL [20] | 44.32 | 39.67 | 66.49 | 88.10 | 17.62 | 82.09 | 42.27 | 37.57 | 52.27 |
| Ours | **45.77** | **39.79** | **73.36** | **89.57** | **17.99** | **82.57** | **42.95** | **45.56** | **54.69** |

Table 8. The VQA performance (%) on the standard test set of TDIUC under the CLT-VQA setting in the random scenario. The memory size in the replay-based methods is 5,000. The best results are highlighted in bold.

| Method | Object_R | Color | Counting | Attribute | Scene_R | Sport_R | Activity_R | Positional_R | AP |
|---|---|---|---|---|---|---|---|---|---|
| Joint | 87.08 | 58.67 | 50.21 | 46.97 | 92.21 | 95.49 | 49.79 | 29.84 | 63.78 |
| Vanilla | 2.17 | 1.19 | 0 | 0 | 0 | 0 | 0 | 16.47 | 2.48 |
| LDAM [5] | 0.28 | 13.39 | 0 | 0.07 | 0 | 0 | 0 | 16.40 | 3.77 |
| GCL [12] | 0.26 | 13.53 | 0 | 0.09 | 0 | 0 | 0 | 16.94 | 3.85 |
| EWC [11] | 57.31 | 23.98 | 22.63 | 0.70 | 37.23 | 45.23 | 23.07 | 14.19 | 28.04 |
| MAS [2] | 56.79 | 27.47 | 38.42 | 0.82 | 31.22 | 45.89 | 24.86 | 12.06 | 29.69 |
| DER [4] | 53.39 | 32.26 | 38.65 | 16.95 | 85.72 | 51.46 | 19.08 | 10.57 | 38.51 |
| ER [6] | 63.6 | 40.39 | 39.37 | 34.57 | 86.81 | 77.31 | 33.55 | 14.96 | 48.82 |
| CVS [18] | 60.1 | 35.46 | 37.74 | 32.96 | 85.23 | 75.45 | 40.74 | 12.51 | 47.52 |
| VQACL [20] | 72.63 | 45.68 | 40.39 | 40.02 | 88.38 | 79.30 | 40.04 | 21.88 | 53.54 |
| Ours | **75.96** | **45.98** | **40.59** | **40.37** | **89.31** | **81.14** | **41.47** | **22.45** | **54.66** |

# F. Additional Experimental Results

In this section, we conduct a comprehensive evaluation of our model across multiple aspects. First, we assess its performance on individual sub-tasks of VQA v2 [8] and TDIUC [10], as well as the overall Average Performance (AP) across all tasks (Sec. F.1). Next, we investigate the impact of the trade-off parameter $\lambda$ in Eq. (6) of the main paper and examine the effect of memory size on model performance (Sec. F.2 and Sec. F.3). Furthermore, we evaluate our approach on a more challenging dataset (VQA-CP v2 dataset [1]) to assess its robustness and generalizability (Sec. F.4). Finally, in Sec. F.5, we compare our method against two continual learning baselines that incorporate long-tailed strategies by integrating the long-tailed methods into our baseline. This comparison provides further insights into the effectiveness of our proposed approach.

## F.1. Fine-grained Results for CLT-VQA

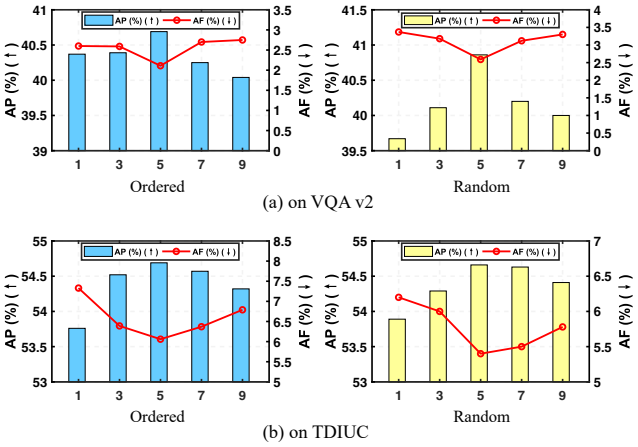
In this section, we present fine-grained results across individual sub-tasks on the VQA v2 [8] and TDIUC [10] datasets. Specifically, Tab. 4, Tab. 5, and Tab. 6 report the model's performance on VQA v2 under the CLT-VQA setting. These results cover 9 tasks in both the ordered and random scenarios, as well as 10 tasks in the random scenario, respectively. Similarly, Tab. 7 and Tab. 8 present the results on TDIUC, evaluating performance in both the ordered and random scenarios. Each column in the tables corresponds to the model's final performance on the respective sub-task, while the last column summarizes the AP across all sub-tasks.

Based on the results, we can derive the following conclusions: (1) In the ordered scenario, as shown in Tab. 4 and Tab. 7, our approach consistently outperforms all competing methods across all tasks, demonstrating substantial im-

provements. Specifically, compared to replay-based methods that more similar to our approach (ER [6], DER [4], CVS [18]), and VQACL [20]), our method achieves an overall AP improvement ranging from 2.71% to 11.2% on VQA v2 [8] and 2.42% to 13.45% on TDIUC [10]. A similar trend is observed in the random scenario, as presented in Tables 5 and 8, further validating the robustness of our method. These results highlight the effectiveness of our approach in preserving strong discriminative capabilities across classes in long-tailed VQA datasets while simultaneously mitigating catastrophic forgetting caused by feature drift in the continual learning paradigm. (2) Tab. 6 presents a more comprehensive evaluation of our method across 10 tasks in the VQA v2 dataset [8]. As observed, our approach consistently achieves the highest performance across all tasks, further demonstrating its effectiveness in mitigating catastrophic forgetting while effectively addressing the long-tail distribution. These results provide additional validation of the robustness and efficacy of our method in the CLT-VQA setting. (3) The results shown in Tab. 7 and Tab. 8 indicate that LDAM [5] and GCL [12] exhibit subpar performance on the earlier sub-tasks in TDIUC [10]. This can be attributed to the distinct class sets in each subtask, which render TDIUC [10] more challenging compared to VQA v2 [8]. Furthermore, these methods are primarily designed to address long-tailed distribution issues and lack mechanisms to effectively mitigate catastrophic forgetting in continual learning settings. Consequently, their performance on earlier sub-tasks deteriorates significantly. However, the performance improvements compared to the continual learning methods such as EWC [11], MAS [2], DER [4] and CVS [18] on the final sub-task demonstrates that LDAM [5] and GCL [12] remain effective for addressing long-tailed data. In contrast, our method not only handles long-tailed distributions robustly but also effectively mitigates catastrophic forgetting, resulting in superior performance across both earlier and later sub-tasks.

## F.2. Hyperparameter Selection for Trade-off Parameter $\lambda$

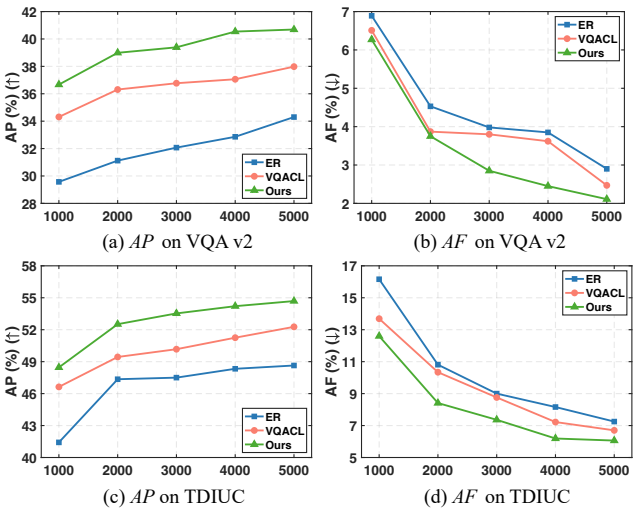
We investigate the influence of an important parameter as defined in Eq. (6) of our main paper. Specifically, we train models with $\lambda \in \{1, 3, 5, 7, 9\}$ in both the ordered and random scenarios for VQA v2 and TDIUC, with the results shown in Fig. 3 (a) and Fig. 3 (b), respectively. As shown in the figure, when $\lambda$ is too small, the $\mathcal{L}_{MFA}$ loss is too weak, which prevents MFA from effectively aligning the updated features with the classifier prototypes. When $\lambda$ increases, the $\mathcal{L}_{MFA}$ loss gradually becomes dominant, which impairs BCP's ability to handle long-tailed VQA data. While $\lambda = 5$, our method achieves a relatively high AP and a lower AF. Therefore, we set $\lambda = 5$ in our experiments.



(a) on VQA v2

(b) on TDIUC

Figure 3. Performance variation with different $\lambda$.



(a) $AP$ on VQA v2

(b) $AF$ on VQA v2

(c) $AP$ on TDIUC

(d) $AF$ on TDIUC

Figure 4. Influence of memory size on VQA v2: (a) $AP$, (b) $AF$; and TDIUC: (c) $AP$, (d) $AF$.

## F.3. Hyperparameter Selection for Memory Size $M$

Fig. 4 show the model performance across different memory sizes on VQA v2 and TDIUC dataset in the ordered scenario, respectively. From these results, we observe that, compared to the ER [6] and VQACL [20], our method consistently achieves the best performance, regardless of the number of stored examples. This demonstrates the effectiveness of our proposed approach for continual long-tailed VQA. Additionally, when the memory size increases, all continual learning methods show clear performance improvements, indicating that a larger memory capacity helps mitigate the forgetting problem. To balance model efficiency and performance, we set the memory size to $M = 5000$ in our experiments.

## F.4. Experiments on Additional Datasets

In addition to reorganizing two widely used VQA datasets, VQA v2 [8] and TDIUC [10], based on question types

Table 9. Results on VQA-CP v2 in the setting of CLT-VQA. *M:* memory size; Ordered: Ordered Scenario; Random: Random Scenario; AP: Final Average Performance (%); AF: Average Forgetting (%). The best results are highlighted in bold.

| Methods | *M* | VQA-CP v2 | | | |
|---|---|---|---|---|---|
| | | Ordered | | Random | |
| | | AP(↑) | AF(↓) | AP(↑) | AF(↓) |
| Joint | - | 32.44 | - | 32.44 | - |
| Vanilla | 0 | 12.32 | 22.34 | 11.68 | 23.9 |
| LDAM [5] | 0 | 13.22 | 17.73 | 13.14 | 17.8 |
| GCL [12] | 0 | 13.54 | 16.64 | 13.46 | 16.21 |
| EWC [11] | 0 | 18.03 | 8.71 | 18.25 | 8.57 |
| MAS [2] | 0 | 19.24 | 7.43 | 19.22 | 8.12 |
| DER [4] | 5000 | 21.34 | 6.36 | 21.61 | 6.55 |
| ER [6] | 5000 | 22.55 | 5.56 | 22.39 | 5.72 |
| CVS [18] | 5000 | 20.66 | 6.99 | 20.97 | 6.82 |
| VQACL [20] | 5000 | 24.05 | 4.82 | 24.14 | 3.93 |
| Ours | 5000 | **25.46** | **3.35** | **25.36** | **3.69** |

Table 10. Comparison of our method with the continual learning baseline enhanced with long-tailed strategies on VQA v2. The best results are highlighted in bold.

| Method | Ordered | | Random | |
|---|---|---|---|---|
| | AP(↑) | AF(↓) | AP(↑) | AF(↓) |
| Baseline | 34.30 | 2.90 | 33.37 | 3.48 |
| +LDAM [5] | 35.33 | 2.86 | 35.13 | 3.43 |
| +GCL [12] | 36.64 | 2.70 | 36.59 | 3.24 |
| +Ours | **40.69** | **2.11** | **40.86** | **2.59** |

Table 11. Comparison of our method with the continual learning baseline enhanced with long-tailed strategies on TDIUC. The best results are highlighted in bold.

| Method | Ordered | | Random | |
|---|---|---|---|---|
| | AP(↑) | AF(↓) | AP(↑) | AF(↓) |
| Baseline | 48.64 | 7.25 | 48.82 | 6.46 |
| +LDAM [5] | 49.57 | 7.21 | 49.66 | 6.36 |
| +GCL [12] | 50.82 | 7.04 | 50.77 | 6.29 |
| +Ours | **54.69** | **6.06** | **54.66** | **5.40** |

to create a sequence of training tasks, we also evaluate our method on the VQA-CP v2 dataset [1], an out-of-distribution (OOD) benchmark. The VQA-CP v2 dataset is constructed by reconfiguring the training and validation splits of the VQA v2 dataset, such that the distribution of answers for each question type differs between the training and test sets. This modification enables the assessment of the model's robustness and its generalization capacity under OOD conditions. From Tab. 9, we can draw the following conclusions: (1) All methods experience a performance drop when facing the OOD scenario, which demonstrates that OOD conditions further exacerbate the challenges in CLT-VQA. (2) Our method consistently achieves the highest AP and the lowest AF in this challenging scenario, demonstrating its robustness and generalizability under OOD conditions.

### F.5. Experiments on Additional Methods

To further validate the effectiveness of our proposed approach, we compare it against two continual learning baselines enhanced with long-tailed strategies by integrating the long-tailed methods LDAM [5] and GCL [12] into our baseline. Tab. 10 and Tab. 11 present the results in the ordered and random scenarios on VQA v2 [8] and TDIUC [10], respectively. From the tables, we can observe that the long-tail methods, by using a fixed-size memory, effectively alleviate catastrophic forgetting and show improvements over the Baseline. However, they still fail to adequately address prototype drift caused by the long-tailed distribution, as well as feature drift of old classes during continual learning. In contrast, our method not only maintains strong discriminative power across classes when dealing with long-tailed VQA data through balanced prototypes but also aligns the drifted features of old classes with the balanced prototypes, effectively mitigating catastrophic forgetting.

## 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 *CVPR*, pages 4971–4980, 2018. 1, 8, 10

[2] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *ECCV*, pages 139–154, 2018. 1, 6, 7, 8, 9, 10

[3] Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *NeurIPS*, pages 2200–2211, 2017. 5

[4] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *NeurIPS*, 2020. 1, 6, 7, 8, 9, 10

[5] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019. 1, 6, 7, 8, 9, 10

[6] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, P Dokania, P Torr, and M Ranzato. Continual learning with tiny episodic memories. In *MTLRL*, 2019. 1, 6, 7, 8, 9, 10

[7] Lenaic Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster wasserstein distance estimation with the sinkhorn divergence. In *NeurIPS*, pages 2257–2269, 2020. 5

[8] 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 *CVPR*, pages 6904–6913, 2017. 1, 6, 8, 9, 10

[9] Viet Huynh, He Zhao, and Dinh Phung. Otlda: A geometry-aware optimal transport approach for topic modeling. In *NeurIPS*, pages 18573–18582, 2020. 3

[10] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *ICCV*, pages 1965–1973, 2017. 1, 2, 6, 8, 9, 10

[11] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *NAS*, 114(13):3521–3526, 2017. 1, 6, 7, 8, 9, 10

[12] Mengke Li, Yiu-ming Cheung, and Yang Lu. Long-tailed visual recognition via gaussian clouded logit adjustment. In *CVPR*, pages 6929–6938, 2022. 1, 6, 7, 8, 9, 10

[13] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *NAS*, 117(40):24652–24663, 2020. 3

[14] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. 3

[15] Pascanu Razvan and Bengio Yoshua. Revisiting natural gradient for deep networks. *arXiv preprint arXiv:1301.3584*, 2013. 6

[16] Jeffrey S Vitter. Random sampling with a reservoir. *TOMS*, 11(1):37–57, 1985. 6

[17] Vy Vo, Trung Le, Tung-Long Vuong, He Zhao, Edwin Bonilla, and Dinh Phung. Parameter estimation in dags from incomplete data via optimal transport. In *ICML*, pages 2700–2725, 2024. 3

[18] Timmy ST Wan, Jun-Cheng Chen, Tzer-Yi Wu, and Chu-Song Chen. Continual learning for visual search with backward consistent feature embedding. In *CVPR*, pages 16702–16711, 2022. 1, 6, 7, 8, 9, 10

[19] Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? In *NeurIPS*, 2022. 3

[20] Xi Zhang, Feifei Zhang, and Changsheng Xu. Vqacl: A novel visual question answering continual learning setting. In *CVPR*, pages 19102–19112, 2023. 1, 6, 7, 8, 9, 10