Towards Robust Visual Question Answering via Prompt-Driven Geometric Harmonization

Yishu Liu[1*], Jiawei Zhu[2*], Congcong Wen[3], Guangming Lu[1†], Hui Lin[3], Bingzhi Chen[2†]

[1]Harbin Institute of Technology Shenzhen, Shenzhen, China
[2]Beijing Institute of Technology, Zhuhai, China
[3]China Academic of Electronics and Information Technology, Beijing, China
liuyishu@stu.hit.edu.cn, luguangm@hit.edu.cn, chenbingzhi@bit.edu.cn

Abstract

Visual Question Answering (VQA) has garnered significant attention as a crucial link between vision and language, aimed at generating accurate responses to visual queries. However, current VQA models still struggle with the challenges of *minority class collapse* and *spurious semantic correlations* posed by language bias and imbalanced distributions. To address these challenges, this paper proposes a novel Prompt-Driven Geometric Harmonization (PDGH) paradigm, which integrates both geometric structure and information entropy principles to enhance the ability of VQA models to generalize effectively across diverse scenarios. Specifically, our PDGH approach is meticulously designed to generate image-generated prompts that are guided by specific question cues, facilitating a more accurate and context-aware understanding of the visual content. Moreover, we project the prompt-visual-question and visual-question joint representations into a unified hypersphere space, applying feature weight self-orthogonality and prompt-information entropy correction constraints to optimize the margin, further alleviating minority class collapse and correcting language bias. To maintain the geometric integrity of the representation space, we introduce multi-space geometric contrast constraints to minimize the impact of spurious priors introduced during training. Finally, a semantic matrix is constructed for the coordinated joint representation to ensure that the learned instances are semantically consistent and improve reasoning ability. Extensive experiments on various general and medical VQA datasets demonstrate the consistent superiority of our PDGH approach over existing state-of-the-art baselines.

Introduction

Humans naturally possess the ability to learn multimodal knowledge asymptotically and excel at using visual cues to think and make wise decisions when confronted with problems. Visual Question Answering (VQA) models aim to bridge the gap between vision and natural language, emulating human reasoning to transition from understanding the semantics of a problem to solving practical issues. Although significant progress has been made in VQA models that mimic human question-answering abilities (Han et al. 2021) and perform well on in-distribution datasets, research has shown that their performance significantly declines on out-of-distribution (OOD) data. In the real world, natural data often follows an unbalanced distribution (Zhang et al. 2023). The primary challenge is to prevent VQA models from relying on language pseudo-priors in the training data and neglecting the association between visual cues and answers, which contradicts the fundamental purpose of visual question answering. Therefore, the ongoing challenge in the field is to develop VQA models that can effectively resist the lure of these pseudo-priors, ensuring that they maintain the integrity of the visual-language association even in the presence of unbalanced data distributions.

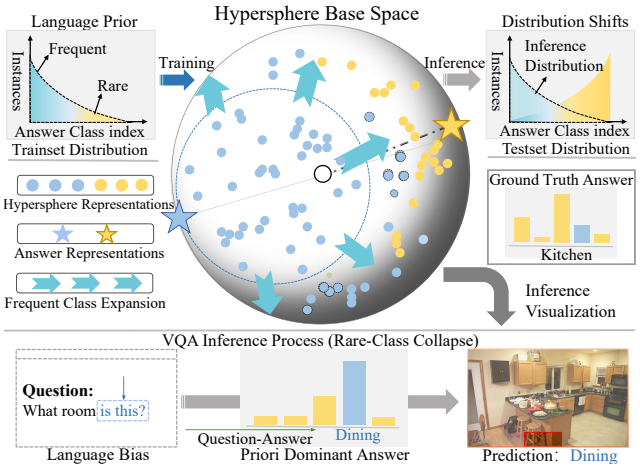


Figure 1: Illustration of the challenges of language bias and imbalanced distributions. Due to inherent differences in class distribution, the unevenness of the class space resulting from the frequent recurrence of common classes and the scarcity of rare answer classes can reduce the robustness of the learned representations in semantic reasoning.

To bridge the gap between human intelligence and neural networks, numerous scholars leveraged bias models during training (Han et al. 2023; Cho et al. 2023) or incorporating question branches (Cadene et al. 2019; Clark,

*Both authors contributed equally to this research.
†Corresponding authors: Bingzhi Chen and Guangming Lu.

Yatskar, and Zettlemoyer 2019), which focused on capturing and recelebrating the potential biases present within each modality or dataset for debiased learning. Additionally, some advanced approaches utilized data augmentation strategies (Liang et al. 2020; Chen et al. 2020; Chen, Zheng, and Xiao 2022) to balance the influence of language priors in data distribution. Existing studies indicate that merely addressing biases in the data to eliminate correlations does not necessarily result in optimal robust learning (Guo et al. 2021), and these methods frequently overlook the complexity of visual question instances hidden in the feature space, which is essential for eliminating implicit spurious correlations. Consequently, representations learned under extreme training conditions may deviate from the true semantics of the instances, leading to the dominance of frequent answer classes during training and the collapse of intrinsic factors associated with minority answer classes (Zhou et al. 2023).

Despite the substantial advancements, current researches still suffer from two critical challenges, i.e., **minority class collapse** and **spurious semantic correlations**, especially under conditions of language bias and imbalanced distributions, where frequently occurring answer classes become overly associated with specific question types. Inspired by the concepts of neural collapse and geometric structure (Zhou et al. 2023; Liu et al. 2023), our research is dedicated to capturing the subtle differences in the representations of instances to enhance the discriminability of instances. As illustrated in Figure 1, we strategically employ manifold contrastive constraints to maximize the angular distance between classes, thereby ensuring sufficient space for sparse answer class data. By refining the spatial arrangement of instance representations, this method addresses the inherent challenges of imbalanced learning and contributes to more accurate and reliable model predictions.

To address these challenges, this paper proposes a robust **P**rompt-**D**riven **G**eometry **H**armonization (PDGH) framework to handle semantic ambiguity and mitigate inference preference risks by promoting uniformity on hyperspherical manifolds and incorporating geometric contrast constraints. Specifically, our PDGH approach initially generates prompt-guided joint representations by fusing vision-question embeddings and prompt-visual-question embeddings. These representations are then used to construct a unit hypersphere space, with margin regularization applied to maintain structural integrity. Moreover, we quantize the energy associated with each instance to capture semantic differences within this hypersphere. By asymptotically minimizing this energy, we establish a uniform point configuration that adheres to principles of spatial orthogonality. This configuration is further refined through prompt-guided semantic calibration, aimed at eliminating any false priors that may distort the learning process. Meanwhile, the PDGH framework dynamically utilizes answer priors and implements multi-space geometric constraints. These constraints are crucial for improving answer-class discrimination, which is achieved by compressing the expansion of frequent class spaces through advanced clustering techniques. Furthermore, we construct a semantic matrix that coordinates these representations, ensuring that the learned instance representations across different spaces are semantically consistent. Our main contributions are summarized as follows:

- To the best of our knowledge, this is the first attempt to explore robust VQA using geometric intuition by applying geometric constraints and minimizing hypersphere energy to construct coordinated feature spaces, enhancing generalization in imbalanced training environments.
- By dynamically coordinating the feature spaces of frequent and rare answer classes, a fine-grained prompt space calibration is designed to leverage the principles of self-orthogonality and semantic information entropy to maintain a balanced learning process.
- Benefiting from well-structured multi-space geometric constraints within the geometric framework, our approach effectively promotes the clustering of instance representations, thus improving the learning effectiveness of underrepresented categories.
- Extensive experiments conducted on multiple biased benchmark datasets, covering a wide range of natural and medical scenarios, demonstrate the effectiveness of the proposed PDGH in achieving robust VQA.

Related Work

VQA Debias Learning

With the advancement of the robust VQA field, researchers have increasingly focused on addressing the challenge of OOD data. To evaluate the generalization capabilities of VQA models, various OOD datasets have been developed according to different protocols. Notably, (Agrawal et al. 2018) introduced the VQA-CP dataset, which utilized an OOD setting based on the original VQA v2 dataset (Goyal et al. 2017). This protocol, characterized by its diverse answer distributions, has become a highly popular benchmark for assessing OOD performance. Additionally, VQA-CE (Dancette et al. 2021) introduced a novel evaluation protocol for VQA v2, designed to diagnose the extent to which VQA models rely on shortcuts. Recently, GQA-OOD (Kervadec et al. 2021) was proposed to address more real-world scenarios by offering a new generalization metric and establishing a larger test benchmark for OOD settings.

VQA Hypersphere Learning

Given the superiority of hypersphere space in robust learning, it is crucial to study its optimization process both during and after training. Inspired by several geometric properties of deep neural networks (Zhou et al. 2023), deep feature spaces are trained via a hypersphere energy minimization scheme. These methods encourage the distribution of weights in feature space to exhibit certain geometric symmetries, facilitating the use of effective regularization techniques, such as the orthonormal (Yang et al. 2020) and orthogonality (Ahmed, Kukleva, and Schiele 2024) methods. In addition, hypersphere learning has demonstrated its benefits in VQA implementations (Basu, Addepalli, and Babu 2023; Guo et al. 2021). Generally, the hardness of a sample is calculated based on the angular distance between the sample and the target class, typically regarded as the class's

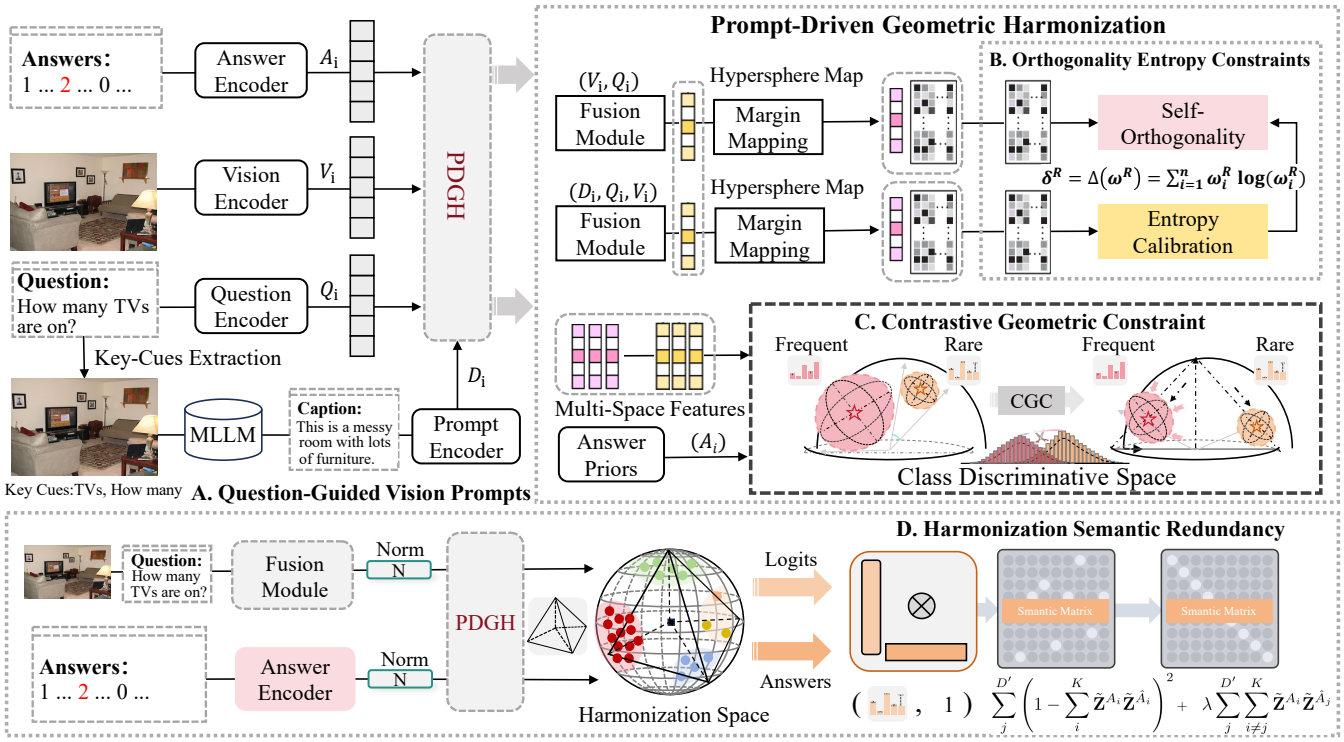


Figure 2: Illustration of our Prompt-Driven Geometry Harmonization (PDGH) framework for combating the challenge of minority class collapse and spurious semantic correlations posed by language bias and imbalanced distributions.

weight vector. Under imbalanced conditions, it is advantageous to mitigate the uncontrolled expansion of frequent classes by enforcing weight symmetries.

Geometry Harmonization Learning

In VQA tasks, imposing geometric structure constraints on answer representations can significantly improve interpretability and generalization. By enforcing these constraints, this approach ensures that intra-class compactness surpasses inter-class separability while maintaining a uniform distribution, to achieve robust generalization under imbalanced conditions. Recently, geometric structure methods have been widely employed in object detection, class-incremental learning, and multimodal representation learning (Poklukar et al. 2022; Hersche et al. 2022; Zhou et al. 2022; Hersche et al. 2022). Relevant to our work, GMC (Poklukar et al. 2022) proposed a geometric multimodal contrast method that promotes robust multimodal representations through geometric alignment. Moreover, NC-FSCIL (Yang et al. 2022) utilized a point regression loss on class features and classifier prototypes to enhance discriminability through a simplex equiangular framework.

Methodology

Problem Statement

Technically, the purpose of our PDGH approach is to explore the semantic reasoning ability of VQA under training-test conditions with two sets of different answer distributions. Given a batch of data samples $\mathcal{B} = \{(\mathcal{V}_i, \mathcal{Q}_i), \mathcal{A}_i\}_{i=1}^{N}$, where $(\mathcal{V}_i, \mathcal{Q}_i)$ is the i-th image-question pairs of samples with the corresponding ground-truth answer $\mathcal{A}_i$, and N is the number of samples. In general, the objective of VQA models is to learn a mapping function $\mathcal{M}_{\text{vqa}}$ to generate a joint representation $\mathcal{R}$. Therefore, the inference of the VQA model can be formulated as follows,

$$\mathcal{P}_i = \mathcal{M}_{\text{vqa}}(\mathcal{V}_i, \mathcal{Q}_i; \theta) = f_\theta\left(e^v(\mathcal{V}_i), e^q(\mathcal{Q}_i)\right), \quad (1)$$

where $\mathcal{P}_i$ denotes probability of answers of the i-th instance, $f_\theta(\cdot)$ is the joint network with parameters θ_m. In addition, the model makes predictions based on the image-question pairs, where $e^v(\cdot)$ is the image encoder, and $e^q(\cdot)$ is the question encoder. As such, the objective function is depicted as:

$$\hat{\mathcal{A}} = \arg\max\left(\mathcal{P}_i\right), \quad (2)$$

where $\hat{\mathcal{A}}$ represents the predicted answer. Note that each instance may have multiple correct answers. Hence, the optimization objective of the training process can be written as:

$$\mathcal{L}\left(\hat{\mathcal{A}}_i, \mathcal{A}\right) = -\frac{1}{B}\sum_{i=1}^{B}\sum_{k=1}^{|\mathcal{A}|}\mathcal{S}_{(i,k)}\log\left(P_i\right), \quad (3)$$

where $\mathcal{S}_{(i,k)}$ is the score of the i-th instance corresponding to the k-th answer of the answer candidates. Notably, Figure 2 illustrates the detailed pipeline of PDGH, which comprises four essential modules, including (a) Question-Guided Vision Prompts, (b) Orthogonality Entropy Constraints, (c) Contrastive Geometric Constraint, and (d) Harmonization Semantic Redundancy.

Question-Guided Vision Prompts

To establish a caption bridge for visual-question interaction, inspired by (Özdemir and Akagündüz 2024; Wang et al. 2024), we build question-guided vision prompts to generate visual captions from images by using key hints from the questions as auxiliary information. The key objective is to create image captions with strong visual representation capabilities. To preserve crucial visual details, it is essential that the generated captions exhibit high accuracy, information consistency, and diversity. Benefiting from the rapid development of multimodal large language models (MLLMs), we use KeyBERT (Grootendorst 2020) to extract keywords from the query, which can be formulated as follows,

$$\left\{\mathcal{K}_i = \left(\mathcal{W}_{i,1}^k, \mathcal{W}_{i,2}^k, \ldots, \mathcal{W}_{i,\mathcal{L}_k}^k\right)\right\}_{i=1}^{\mathcal{N}_k} = \mathrm{Keybert}(\mathcal{Q}), \quad (4)$$

where $\mathcal{Q}$ represents the query text, $\mathcal{K}_i$ is the i-th key phrase, $\mathcal{L}_k$ is the length of key phrases, and $\mathcal{N}_k$ is the number of extracted key phrases. Next, we use the extracted keywords to guide the generation of image captions, which can be achieved by integrating the keywords into the input of MLLMs such as CogVLM2 (Wang et al. 2023), i.e.,

$$\left\{\mathcal{D}_z^k = \left(\mathcal{W}_{z,1}^{ck}, \mathcal{W}_{z,2}^{ck}, \ldots, \mathcal{W}_{z,L_p}^{ck}\right)\right\}_{z=1}^{\mathcal{N}_c} = \mathrm{CogVLM}(\mathcal{V}, \mathcal{K}), \quad (5)$$

where $\mathcal{V}$ represents the image, $\mathcal{K}$ indicates the generated question keywords, $\mathcal{L}_p$ is the length of the generation of prompt descriptions, and $\mathcal{N}_c$ is the number of extracted prompt description phrases.

For each generated description, we employ a shared question text encoder $e^q\left(\mathcal{Q}_i\right)$ to extract embeddings from each prompt text. To obtain the prompt-guided joint representation, we introduce an additional prompt-question-vision fusion branch $\mathcal{R}^p = f_\theta\left(e^v\left(\mathcal{V}_i\right), e^q\left(\mathcal{Q}_i\right), e^q\left(\mathcal{D}_i\right)\right)$. Notably, the prompt-guided joint representation, which shares the homeomorphic sample structure, is excluded from the reasoning process to ensure fairness.

Hypersphere Mapping Module

Generally, most existing VQA feature space debiasing methods utilize margin loss to construct a unit hypersphere space, which encourages discriminability among instances during training. Recognizing the intrinsic influence of spurious priors, we integrate the geometric structure of PDGH to calibrate the instance space, thereby enhancing the robustness of VQA models under imbalanced distribution conditions. In this framework, the hypersphere space preserves the same sample structure as the original joint representation, and we apply the concept of angular margin to precisely manipulate and refine the feature space.

Motivated by (Guo et al. 2021; Basu, Addepalli, and Babu 2023), our work projects the joint representation $\mathcal{R}$ and $\mathcal{R}^p$ onto a hypersphere with a constant unit radius. Specifically, we initialize the angle margin by applying L_2-normalization to the weight vector $\mathcal{W}_i$ and the joint representations $\mathcal{R}_i$ and $\mathcal{R}_i^p$, ensuring that the posterior probability is determined by the angle θ_i. Let θ_i denote the angle between $\mathcal{R}_i$ and $\mathcal{W}_i$. Therefore, the unregularized and regularized logits of each instance are defined as follows:

$$Logit_i = \mathcal{W}_{\mathbf{i}}^\top \mathcal{R}_i, \quad (6)$$

$$\hat{Logit}_i = \mathcal{W}_{\mathbf{i}}^\top \mathcal{R}_i = \|\mathcal{W}_{\mathbf{i}}\| \, \|\mathcal{R}_i\| \cos\theta_i = s\left(\cos\theta_i\right). \quad (7)$$

During training, we retain the original unregularized logit $Logit_i$. Notably, the instance representations $\mathcal{R}_i$ are L_2-normalized, where $\|\mathcal{W}_i\| = 1, \|\mathcal{R}_i\| = 1$. Here, s represents the radius of the hypersphere, and the instance representations are mapped onto this hypersphere. After normalization, the predictions of the cosine classifier depend solely on the angles between the joint representations and the weight vector. Inspired by (Guo et al. 2021; Basu, Addepalli, and Babu 2023; Liu et al. 2017; Shen et al. 2021), the number of answers in the training process and the difficulty of learning the semantic knowledge contained in different answers have certain differences. Therefore, we incorporate this implicit relationship into the angle margin and use frequency and instance difficulty to optimize the hypersphere space between different answers, that is,

$$\hat{Logit}_i = s \cdot \cos\left(\theta_i + m[i]\right), \quad (8)$$

Where $m[i]$ is the adaptive instance angle margin, and the angle margin $(\theta_i + m[i])$ in the representation R_i enables models to distinguish between frequent/rare instances.

Orthogonality Entropy Constraints

The feature space evolves dynamically during training, with frequently occurring instances actively expanding and compressing the space of rare answer classes, which leads to increased congestion in the instance feature space.

Self-Orthogonality Constraints To maintain the feature space for minority class question-answering instances and avoid passive collapse, we impose a self-orthogonality constraint on the hypersphere space to maximize the angular separation between these instance weight vectors, thereby enhancing their mutual orthogonality. Inspired by the Thomson problem (Liu et al. 2018, 2023), which seeks to distribute N electrons as uniformly as possible on a unit sphere with minimal potential energy, we regulate the hypersphere energy functional of each instance's weights accordingly:

$$\arg\min_{\boldsymbol{W}_N} \boldsymbol{E}_0 = \arg\max_{\boldsymbol{W}_N} \prod_{i \neq j} \left\|\hat{\boldsymbol{w}}_i - \hat{\boldsymbol{w}}_j\right\|, \quad (9)$$

where $\hat{\boldsymbol{w}}_i$ and $\hat{\boldsymbol{w}}_j$ represent the i-th and j-th instance representation weights, and $\boldsymbol{W}_N$ denotes the set of all instance weight vectors. To further refine the energy function, we define the following equation:

$$\begin{aligned} \boldsymbol{E}_{s,d}\left(\hat{\boldsymbol{W}}_i\Big|_{i=1}^N\right) &= \sum_{i=1}^N \sum_{j=1, j\neq i}^N K_s\left(\|\hat{\boldsymbol{w}}_i - \hat{\boldsymbol{w}}_j\|\right) \\ &= \begin{cases} \sum_{i\neq j} \|\hat{\boldsymbol{w}}_i - \hat{\boldsymbol{w}}_j\|^{-s}, s > 0, \\ \sum_{i\neq j} \log\left(\|\hat{\boldsymbol{w}}_i - \hat{\boldsymbol{w}}_j\|^{-1}\right), s = 0. \end{cases} \end{aligned} \quad (10)$$

where $K_s\left(\hat{\boldsymbol{w}}_i, \hat{\boldsymbol{w}}_j\right)$ denotes a Riesz kernel function (Riesz 1909) and $\hat{\boldsymbol{w}}_i = \frac{\boldsymbol{w}_i}{\|\boldsymbol{w}_i\|}$ is the weight of the i-th instance

representation projected onto the unit hypersphere $\mathbb{S}^d = \left\{R_i \in \mathbb{R}^{d+1} \mid \|R\| = 1\right\}$. The minimizer of this pairwise energy sum asymptotically corresponds to a uniform distribution on a hypersphere. In addition, the kernel function is parameterized by s, which dynamically adjusts the relative similarity positions of the weights of different answer categories to distribute the feature points on the hypersphere space. Moreover, $\mathcal{L}_{\mathrm{SOC}} = \frac{1}{N}\sum_{i\neq j}\frac{1}{\|\boldsymbol{w}_i - \boldsymbol{w}_j\|^2+\epsilon}$ minimizes the hyperspherical energy of class weight using gradient descent during back-propagation, where ϵ is a small positive constant introduced to prevent division by zero.

Entropy-Calibration Constraints In order to consider the consistency of the prompt information entropy and the joint embedding entropy, we explore an entropy-calibration constraint to correct the feature distribution. By adjusting the prompt-driven representation entropy δ_p^R, we innovatively optimize the semantically consistent representation distribution to fuse the intuitive semantics derived from visual cues and the questions, which can be defined as:

$$\delta^R = \Delta\left(\boldsymbol{w}^R\right) = \sum_{i=1}^{n} \boldsymbol{w}_i^R \log\left(\boldsymbol{w}_i^R\right). \tag{11}$$

To narrow the gap between joint representation and prompt-driven representation, we propose orthogonality entropy constraints $\mathcal{L}_{\mathrm{OEC}}$ to correct the information distribution of joint representation. In the hypersphere space, the information entropy δ^R and δ_p^R are calculated to align the information distribution. Specifically, we generate the square difference of adaptive entropy $\mathcal{L}_{\mathrm{OEC}}$ for information consistency:

$$\mathcal{L}_{\mathrm{OEC}} = \sum_{i=1}^{N}\left(\delta^R - \delta_p^R\right)^2 + \mathcal{L}_{\mathrm{SOC}}. \tag{12}$$

Contrastive Geometric Constraint

After constructing the instance uniformity distribution space, another major challenge is to mitigate the over-inflation of frequent classes while ensuring sufficient training space for rare classes. Inspired by (Poklukar et al. 2022; Koishekenov et al. 2023), we design a contrastive geometric constraint to further optimize the feature space. In particular, the learning process is refined by accounting for sample difficulty and separation in the feature space using geometric contrast constraints, which could help to normalize the feature space. These constraints also create a clear distinction between samples with different ground truth answers within the feature space. Considering a sample with index j, the set of positive examples in a mini-batch is defined as:

$$S_{(i,j)}^p = \exp\left(\mathrm{sim}\left(\mathcal{R}_i, \mathcal{R}_j^p\right)/\tau\right), \tag{13}$$

where $S_{(i,j)}^p$ denotes the similarity between the joint representation $\mathcal{R}_i$ and the prompt representation $\mathcal{R}_j^p$ corresponding to the i-th and j-th samples in a mini-batch. For the given instances, we define two representations with the same answer as positive pairs $\left(\mathcal{R}_i, \mathcal{R}_j^p\right)$, where $i = \{1, 2, \ldots, N\}$, and regard the rest as negative pairs.

To measure the distance of the curve space and make the optimization process of the latent space conform to geometric intuition. Given two instance representations $\mathcal{R}_i$ and $\mathcal{R}_j^p$ on the unit sphere, the spherically symmetric hypersphere has a metric that defines the geodesic distance, i.e., $sim(\mathcal{R}_i, \mathcal{R}_j^p) = \left\langle \mathcal{R}_i, \mathcal{R}_j^p \right\rangle$, where $\left\langle \mathcal{R}_i, \mathcal{R}_j^p \right\rangle$ is the standard inner product in $\mathcal{R}^n$. Therefore, for two representations $\mathcal{R}_i$ and $\mathcal{R}_j^p$ on the unit sphere, we define the contrastive geometric constraint by using the geodesic distance between them,

$$\mathcal{L}_{\mathrm{CGC}} = \frac{-1}{\left|\mathcal{P}_{(i)}\right|}\sum_{p\in\mathcal{P}_{(i)}} \log\frac{e^{S_{(i)}^p/\tau}}{\sum_p e^{S_{(i)}^p/\tau} + \sum_n R_{(i)}^n e^{S_{(i)}^n/\tau}}, \tag{14}$$

where τ is the temperature parameter. Benefiting from the natural angular geodesic distance, we explore prompt-driven geometric harmonization optimization within the same answer class, thereby enforcing intrinsic geometric constraints on the feature landscape.

Harmonization Semantic Redundancy

In this part, we integrate these components into a prompt-driven geometric co-training framework, which utilizes additional visual prompt knowledge to establish a bridge between questions and visual prompts, thereby creating a coordinated feature space. Then, a geometrically optimized representation is achieved for prediction, utilizing the integrated logits $comb_i = \frac{\hat{Logit}_i + Logit_i}{2}$ for the final answer prediction $\hat{\mathcal{A}}_i^{comb}$. Hence, we can minimize the loss between the predicted answer and the ground truth label, that is,

$$\mathcal{L}_{\mathrm{VQA}} = \sum_{i=1}^{|\mathcal{A}|} -\hat{\mathcal{A}}_i^{comb} \log\frac{\exp\left(s\cdot\cos\left(\theta_i + m[i]\right)\right)}{\sum_{j=1}^{|\mathcal{A}|}\exp\left(s\cdot\cos\left(\theta_j + m[j]\right)\right)}. \tag{15}$$

In addition, we propose a harmonization semantic redundancy (HSR) module to further alleviate information redundancy and obtain a more coherent semantic representation. Specifically, we first normalize the joint prediction $\hat{A}_i$ and answer representation A_i along the dimension of the batch B and then calculate their cross-correlation Z. Therefore, the decorrelation between semantic constraints and feature dimensions can be formulated as follows:

$$\mathcal{L}_{\mathrm{HR}} = \frac{1}{D_1}\{\underbrace{\sum_j^{D'}\left(1 - \sum_i^K \tilde{\mathbf{Z}}^{A_i}\tilde{\mathbf{Z}}^{\hat{A}_i}\right)^2}_{\text{Semantic invariance}} + \underbrace{\lambda\sum_j^{D'}\sum_{i\neq j}^K \tilde{\mathbf{Z}}^{A_i}\tilde{\mathbf{Z}}^{\hat{A}_j}}_{\text{Semantic gap reduction}}\}. \tag{16}$$

Hence, the objective of the HSR module can be defined as:

$$\mathcal{L}_{\mathrm{HSR}} = \mathcal{L}_{\mathrm{VQA}} + \mathcal{L}_{\mathrm{HR}}. \tag{17}$$

Training and Optimization

Based on the above analyses, the overall training objective of the proposed PDGH approach is a combination of multiple objective functions from different modules, i.e.,

$$\mathcal{L}_{\mathrm{Total}} = \mathcal{L}_{\mathrm{OEC}} + \mathcal{L}_{\mathrm{CGC}} + \mathcal{L}_{\mathrm{HSR}}. \tag{18}$$

As a result, these components are integrated into a prompt-driven geometric harmonization framework that operates cohesively, enabling each component to mutually benefit from and enhance the performance of the others.

Datasets		VQA-CP v2				VQA-CP v1			
Methods / Reference		Overall	Y/N	Num	Others	Overall	Y/N	Num	Others
UpDn (Anderson et al. 2018)	CVPR'18	39.74	42.27	11.93	46.05	37.96	42.79	12.41	42.53
AdvReg (Ramakrishnan, Agrawal, and Lee 2018)	NeurIPS'18	41.17	65.49	15.48	35.48	43.43	74.16	12.44	25.32
LMH (Clark, Yatskar, and Zettlemoyer 2019)	EMNLP'19	52.15	70.29	44.10	44.86	55.73	78.59	24.68	45.47
GGE-iter (Han et al. 2021)	ICCV'21	57.12	87.35	26.16	49.77	59.82	85.52	28.93	46.67
COB (Jha et al. 2023)	WACV'23	57.53	88.36	28.81	49.27	60.98	87.41	32.02	46.34
GENB (Cho et al. 2023)	CVPR'23	59.15	88.03	40.05	49.25	62.74	86.18	43.85	**47.03**
GGD (Han et al. 2023)	TPAMI'23	59.37	88.23	38.11	49.82	-	-	-	-
PWVQA (Vosoughi et al. 2024)	TMM'24	59.06	88.26	52.89	45.45	-	-	-	-
CVIV (Pan et al. 2024)	TMM'24	60.08	88.85	40.77	50.30	-	-	-	-
PDGH	Ours	**61.68**	**89.29**	**53.13**	**50.32**	**64.56**	**89.56**	**47.35**	46.01

Table 1: Comparisons with the state-of-the-art baselines on the VQA-CP v2 and VQA-CP v1 datasets.

Datasets	VQA-CE			GQA-OOD		
Methods	Overall	Counter	Easy	All	Tail	Head
CSS	53.55	34.36	62.08	44.24	41.20	46.11
GENB	57.87	34.80	68.15	49.43	**45.63**	51.76
RMLVQA	58.05	35.01	68.21	49.07	44.50	51.88
CVIV	-	36.12	-	49.36	-	-
PDGH	**59.10**	**36.21**	**68.31**	**49.56**	44.93	**52.01**

Table 2: Comparisons on VQA-CE and GQA-OOD datasets.

Datasets		SLAKE-LB		
Methods / Reference		Overall-CP	Open-CP	Closed-CP
SAN	CVPR'18	33.22	63.30	6.77
GGE	ICCV'21	41.13	73.10	13.02
DeBCF	MICCAI'23	58.69	**78.08**	28.19
RMLVQA	CVPR'23	78.24	65.00	89.89
PDGH	Ours	**81.21**	76.81	**90.53**

Table 3: Comparisons on the SLAKE-LB dataset.

Experiments

Experimental Setup

Dataset and Evaluation Metric In our experiments, we select various out-of-distribution benchmarks to assess the robustness of models against real-world biases, such as VQA-CP v2, VQA-CP v1 (Agrawal et al. 2018), GQA-OOD (Kervadec et al. 2021), and VQA-CE (Dancette et al. 2021). Following VQA-CP (Agrawal et al. 2018), we develop a Semantically-Labeled Knowledge-Enhanced under Language Bias (SLAKE-LB) benchmark based on SLAKE (Liu et al. 2021) to verify the performance of our method in the medical domain. All experiments utilize the standard VQA evaluation metric (Antol et al. 2015).

Implementation Details To evaluate the efficiency and generalization capabilities of our proposed model, we conduct comparative experiments with the most relevant existing methods. In our experiments, we implement the PDGH model on a single RTX 3090 GPU with PyTorch. The AdamW optimizer is used with a weight decay of 0.001, a learning rate of 0.001, and a batch size of 512.

Comparisons with State-of-the-Arts

Evaluation on VQA-CP v2 and VQA-CP v1 As demonstrated in Table 1, we provide a comprehensive analysis of our proposed PDGH method. In particular, we report the overall accuracy and the accuracy for specific question types, including "yes/no", "number", and "other". Compared to the second-best performing method, our method improves the overall accuracy on the VQA-CP v2 and VQA-CP v1 datasets by 1.6% and 1.82%, respectively.

Evaluation on VQA-CE and GQA-OOD To assess general applicability in diverse real-world scenarios, we evaluate the debiasing performance of our method on the GQA-OOD and VQA-CE datasets, as shown in Table 2. Notably, the VQA-CE test set emphasizes false negative examples involving questions, visual information, and answers, requiring special attention to performance on counterfactual datasets. The experimental results reveal that our proposed method effectively mitigates the impact of spurious correlations and enhances generalization, with overall accuracy improvements of 1.05% on VQA-CE and 0.20% on GQA-OOD compared to the second-best models.

Evaluation on SLAKE-LB To evaluate the generalization performance in medical scenarios, we follow the settings in (Zhan et al. 2023) and construct the out-of-distribution medical benchmark, i.e., SLAKE-LB. We re-split this dataset using the same partitioning ratio as VQA-CP v2 to maintain a consistent sample structure. As presented in Table 3, the experimental results demonstrate that our method exhibits superior generalization performances, highlighting its potential for application in auxiliary diagnostic scenarios.

Ablation Study

To explore the individual impact of each component, we perform ablation studies on the VQA-CP v2 dataset. As shown in Table 4, we can conclude that: 1) The combination of QVP and OEC leads to substantial performance

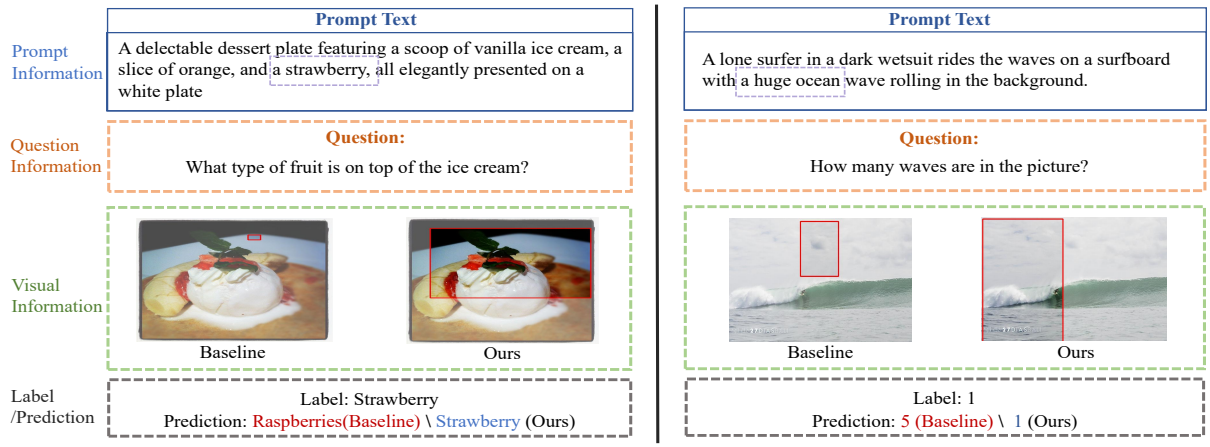


Figure 3: We provide two examples of harmonization learning driven by generative description, showing that the descriptive knowledge from vision can effectively prompt the model to pay attention to certain fine-grained visual cues.

Methods	QVP	OEC	CGC	HSR	Overall-CP
w/ QVP	✓				60.32
w/ CGC			✓		60.76
w/ QVP+OEC	✓	✓			60.69
w/ QVP+OEC+CGC	✓	✓	✓		61.43
PDGH (Ours)	✓	✓	✓	✓	**61.69**

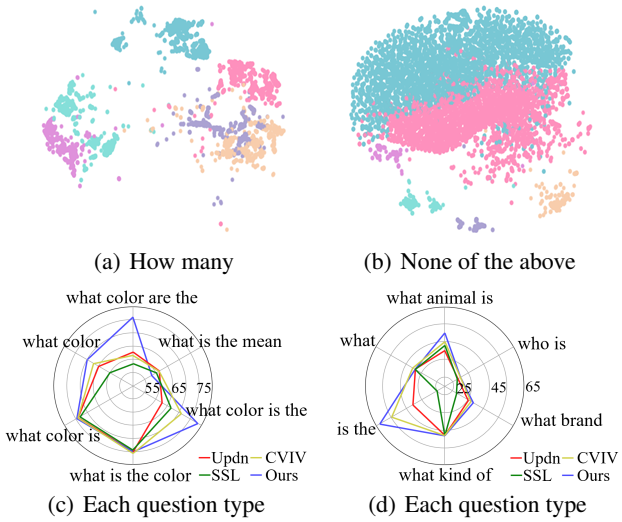
Table 4: Ablation experiments on the VQA-CP v2 dataset.

improvements. This demonstrates the efficacy of integrating prompt generation descriptions with feature space constraints, which enhances both the richness of visual information and the collaborative optimization of the feature space. 2) CGC further optimizes the clustering of feature space and reduces passive class collapse, thus improving the overall representation. 3) The incorporation of HSR improves the final joint evidence, further generating more semantically relevant answer predictions for robust VQA.

Visualization Results

Attention Region Visualization We visualize the generated prompts and attention areas in VQA. As shown in Figure 3, our method generates semantically rich prompts for each question and corresponding image to guide the model during training. For example, in a surfing scenario, the baseline model may focus on irrelevant areas, while our PDGH prioritizes the most important areas guided by geometric learning, effectively capturing rich semantic information.

Visualization Analysis of Question Types As illustrated in Figure 4(a) and Figure 4(b), we visualize the feature distribution of instances across two question types. Notably, our model separates instances with different answers for the same question type, while instances with the same answer remain closely clustered, highlighting the model's strong semantic reasoning ability. To make a fair comparison, we have selected twelve question types and displayed the accuracy of each type in a radar chart in Figure 4(c) and Figure 4(d). It can be observed that our PDGH could achieve the highest performance for most question types, demonstrating the robustness of our method across different questions.



(a) How many (b) None of the above

(c) Each question type (d) Each question type

Figure 4: Visualization of the distribution of answer representations of the same type and the performance of different types by our proposed method on the VQA CP v2 dataset.

Conclusion

This study presented a novel Prompt-Driven Geometric Harmonization framework that integrated geometric constraints and prompt corrections for robust representation learning under imbalanced distributions. By optimizing feature distribution within a hypersphere space, PDGH could enhance discrimination between answer types and mitigate the over-inflation of common classes. Experimental results have demonstrated that PDGH outperforms state-of-the-art methods across multiple benchmark datasets, showcasing exceptional generalization capabilities in real-world applications.

Acknowledgments

This work was supported in part by the Shenzhen Fundamental Research Fund (Grant NO. JCYJ20240813105900002), in part by the NSFC fund (Grant NOs. 62302172, 62176077), in part by the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (Grant NO. 2022B1212010005), in part by the Guangdong International Science and Technology Cooperation Project (Grant NO. 2023A0505050108), and in part by the Shenzhen Key Technical Project (Grant NOs. JSGG20220831092805009, JSGG20220831105603006, JSGG20201103153802006, KJZD20230923115117033).

References

Agrawal, A.; Batra, D.; Parikh, D.; and Kembhavi, A. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4971–4980.

Ahmed, N.; Kukleva, A.; and Schiele, B. 2024. OrCo: Towards better generalization via orthogonality and contrast for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 28762–28771.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6077–6086.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2425–2433.

Basu, A.; Addepalli, S.; and Babu, R. V. 2023. Rmlvqa: A margin loss approach for visual question answering with language biases. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11671–11680.

Cadene, R.; Dancette, C.; Cord, M.; Parikh, D.; et al. 2019. Rubi: Reducing unimodal biases for visual question answering. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Chen, L.; Yan, X.; Xiao, J.; Zhang, H.; Pu, S.; and Zhuang, Y. 2020. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10800–10809.

Chen, L.; Zheng, Y.; and Xiao, J. 2022. Rethinking data augmentation for robust visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 95–112.

Cho, J. W.; Kim, D.-J.; Ryu, H.; and Kweon, I. S. 2023. Generative bias for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11681–11690.

Clark, C.; Yatskar, M.; and Zettlemoyer, L. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Dancette, C.; Cadene, R.; Teney, D.; and Cord, M. 2021. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1574–1583.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6904–6913.

Grootendorst, M. 2020. Keybert: Minimal keyword extraction with bert. *Zenodo*.

Guo, Y.; Nie, L.; Cheng, Z.; Ji, F.; and Zhang, J. 2021. Overcoming Language Priors with Adapted Margin Cosine Loss. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.

Han, X.; Wang, S.; Su, C.; Huang, Q.; and Tian, Q. 2021. Greedy gradient ensemble for robust visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1584–1593.

Han, X.; Wang, S.; Su, C.; Huang, Q.; and Tian, Q. 2023. General greedy de-bias learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(8): 9789–9805.

Hersche, M.; Karunaratne, G.; Cherubini, G.; Benini, L.; Sebastian, A.; and Rahimi, A. 2022. Constrained few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9057–9067.

Jha, A.; Patro, B.; Van Gool, L.; and Tuytelaars, T. 2023. Barlow constrained optimization for visual question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1084–1093.

Kervadec, C.; Antipov, G.; Baccouche, M.; and Wolf, C. 2021. Roses are red, violets are blue... but should vqa expect them to? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2776–2785.

Koishekenov, Y.; Vadgama, S.; Valperga, R.; and Bekkers, E. J. 2023. Geometric contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 206–215.

Liang, Z.; Jiang, W.; Hu, H.; and Zhu, J. 2020. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3285–3292.

Liu, B.; Zhan, L.-M.; Xu, L.; Ma, L.; Yang, Y.; and Wu, X.-M. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 1650–1654.

Liu, W.; Lin, R.; Liu, Z.; Liu, L.; Yu, Z.; Dai, B.; and Song, L. 2018. Learning towards minimum hyperspherical energy. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; and Song, L. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 212–220.

Liu, W.; Yu, L.; Weller, A.; and Schölkopf, B. 2023. Generalizing and decoupling neural collapse via hyperspherical uniformity gap. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Özdemir, Ö.; and Akagündüz, E. 2024. Enhancing Visual Question Answering through Question-Driven Image Captions as Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1562–1571.

Pan, Y.; Liu, J.; Jin, L.; and Li, Z. 2024. Unbiased visual question answering by leveraging instrumental variable. *IEEE Transactions on Multimedia (TMM)*.

Poklukar, P.; Vasco, M.; Yin, H.; Melo, F. S.; Paiva, A.; and Kragic, D. 2022. Geometric multimodal contrastive representation learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 17782–17800.

Ramakrishnan, S.; Agrawal, A.; and Lee, S. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Riesz, F. 1909. *Sur les opérations functionnelles linéaires*. Gauthier-Vllars.

Shen, J.; Xiao, Z.; Zhen, X.; and Zhang, L. 2021. Spherical zero-shot learning. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 32(2): 634–645.

Vosoughi, A.; Deng, S.; Zhang, S.; Tian, Y.; Xu, C.; and Luo, J. 2024. Cross modality bias in visual question answering: A causal view with possible worlds VQA. *IEEE Transactions on Multimedia (TMM)*, 8609–8624.

Wang, B.; Ma, Y.; Li, X.; Gao, J.; Hu, Y.; and Yin, B. 2024. Bridging the Cross-Modality Semantic Gap in Visual Question Answering. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 1–13.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Yang, S.; Deng, W.; Wang, M.; Du, J.; and Hu, J. 2020. Orthogonality loss: Learning discriminative representations for face recognition. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 31(6): 2301–2314.

Yang, Y.; Yuan, H.; Li, X.; Lin, Z.; Torr, P.; and Tao, D. 2022. Neural collapse inspired feature-classifier alignment for few-shot class-incremental learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Zhan, C.; Peng, P.; Zhang, H.; Sun, H.; Shang, C.; Chen, T.; Wang, H.; Wang, G.; and Wang, H. 2023. Debiasing medical visual question answering via counterfactual training. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 382–393.

Zhang, Y.; Kang, B.; Hooi, B.; Yan, S.; and Feng, J. 2023. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(9): 10795–10816.

Zhou, D.-W.; Wang, F.-Y.; Ye, H.-J.; Ma, L.; Pu, S.; and Zhan, D.-C. 2022. Forward compatible few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9046–9056.

Zhou, Z.; Yao, J.; Hong, F.; Zhang, Y.; Han, B.; and Wang, Y. 2023. Combating representation learning disparity with geometric harmonization. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 20394–20408.