BEYOND FEATURE SHIFT: MEASURING AND MITIGATING OUTPUT DISTRIBUTION COLLAPSE IN CROSS-DOMAIN VQA

Anonymous authors
Paper under double-blind review

ABSTRACT

Although vision-language models (VLMs) achieve strong performance on general-domain benchmarks, their effectiveness often degrades substantially when applied to domain-specific visual question answering (VQA) tasks such as medical imaging, remote sensing, and art understanding. Existing studies mainly attribute this degradation to feature-level distribution mismatch and focus on representation alignment. In this work, we show that such explanations are insufficient to characterize the failure mechanisms of pretrained VLMs under domain shift. Through a systematic study across six domain-specific VQA benchmarks and multiple VLM architectures, we identify a previously overlooked phenomenon termed output distribution collapse, characterized by prediction distribution collapse toward dominant answer prototypes, confidence miscalibration, and decision-margin instability. To quantify this phenomenon, we introduce an entropy-based **Collapse Index (CI)** that measures the severity of predictive distribution degeneration. We show that CI consistently exhibits a strong negative correlation with cross-domain VQA accuracy across datasets and architectures. Motivated by these findings, we propose **Distribution Collapse Stabilization (DCS)**, a lightweight output-space learning framework that jointly promotes distribution dispersion, reliability consistency, and decision separability. Extensive experiments on six benchmarks and ten representative VLM backbones demonstrate consistent improvements, yielding an average gain of 4.1% over pretrained baselines while outperforming other adaptation methods. Our findings identify output distribution collapse as a previously underexplored failure mechanism in cross-domain VQA and establish output-space stabilization as a promising alternative to conventional feature-space adaptation.

1 INTRODUCTION

Vision-language models (VLMs) have achieved remarkable progress in visual question answering (VQA) through large-scale multimodal pretraining and instruction tuning Radford et al. (2021); Li et al. (2022); Kim et al. (2021); Li et al. (2023); Liu et al. (2024b;a). However, their performance often degrades substantially when transferred to domain-specific scenarios such as medical diagnosis, remote sensing, and art understanding, where the target distributions differ significantly from pretraining data Wang et al. (2022b); Lobry et al. (2020). Despite extensive studies on cross-domain adaptation, the underlying mechanism responsible for this degradation remains poorly understood.

Existing approaches mainly attribute cross-domain failure to feature-level distribution shift and address it through representation alignment, prompt learning, or parameter-efficient adaptation Zhou et al. (2022); Khattak et al. (2023a;b); Hu et al. (2022). While effective, these methods primarily operate in feature space and overlook how predictive distributions behave under domain shift.

In this work, we revisit cross-domain VQA from the perspective of output-space behavior and identify a previously underexplored phenomenon termed output distribution collapse. Across six VQA benchmarks and multiple vision-language architectures, we observe that predictive distributions progressively collapse toward a small set of dominant answer prototypes under domain shift. Moreover,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

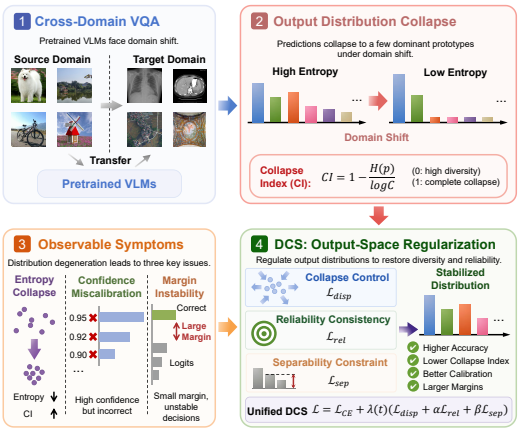


Figure 1: Domain shift induces output distribution collapse, where predictive probabilities collapse toward a few dominant answer prototypes. We propose CI and DCS to restore distribution diversity and improve robustness.

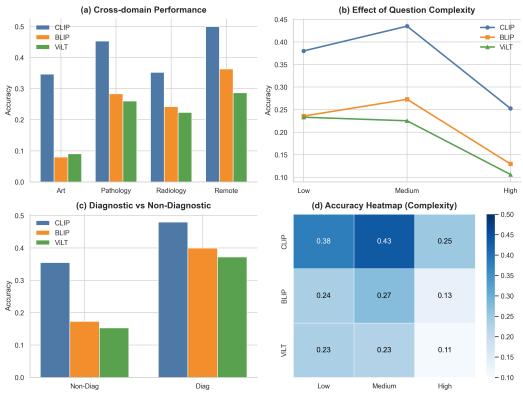


Figure 2: VQA analysis. (a) Accuracy across domains. (b) Performance versus question complexity. (c) Diagnostic vs. non-diagnostic questions. (d) Summary of Complexity-Dependent Performance Trends Across Models.

prediction collapse is accompanied by confidence miscalibration Balanya et al. (2024); Chen et al. (2025) and unstable decision margins Ravikumar et al. (2025).

To quantify this phenomenon, we introduce a normalized **Collapse Index (CI)**, which measures the entropy deficiency of predictive distributions. Across datasets and architectures, lower CI correlates with higher cross-domain VQA accuracy, indicating that output distribution collapse constitutes an important bottleneck for robust generalization. As illustrated in Figure 1, cross-domain VQA failure can be interpreted as a progressive degeneration of predictive distributions. Motivated by these observations, we propose **Distribution Collapse Stabilization (DCS)**, a lightweight output-space learning framework that preserves distribution diversity, confidence reliability, and decision separability without modifying model architectures or introducing additional trainable parameters. Experiments on six cross-domain VQA benchmarks and ten vision-language backbones demonstrate that DCS consistently improves performance by 4.1% on average while maintaining zero parameter overhead. The gains further generalize to recent large vision-language models, including Qwen3-VL Bai et al. (2025) and InternVL3.5 Wang et al. (2025), demonstrating the effectiveness of output-space stabilization across architectures and scales. Due to space limitations, additional related work, detailed analyses of output distribution collapse, theoretical analyses of DCS, comprehensive ablation studies, and supplementary experimental results are provided in the **Appendix**.

2 Understanding Cross-Domain VQA Failures

Existing studies on cross-domain VQA primarily attribute performance degradation to feature-level distribution shift or modality misalignment. However, these explanations cannot fully characterize how predictive behavior deteriorates in pretrained vision-language models, where decoding dynamics and answer priors influence final decisions. To address this gap, we conduct a systematic study on six datasets. We adopt a unified candidate-answer protocol, where each sample is evaluated over a fixed candidate pool containing the ground-truth answer and high-frequency distractors.

Cross-Domain Performance and Task Sensitivity. As shown in Figure 2(a), all models suffer significant performance drops under domain shift, especially in medical and artistic domains. Performance further decreases with increasing question complexity (Figure 2(b)), indicating instability in compositional reasoning. In addition, diagnostic questions exhibit more severe degradation than non-diagnostic ones (Figure 2(c)), suggesting over-reliance on dataset-specific priors rather than grounded multimodal understanding.

Output Distribution Collapse under Domain Shift. We observe a consistent phenomenon of output distribution collapse, where predictions collapse toward a small set of dominant answers.

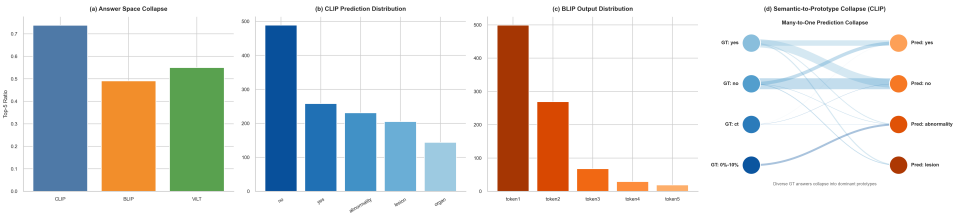


Figure 3: Output distribution collapse. (a) Top-5 concentration ratios. (b-c) Predictions collapse toward dominant answer prototypes. (d) Many-to-one mapping from ground-truth to predictions.

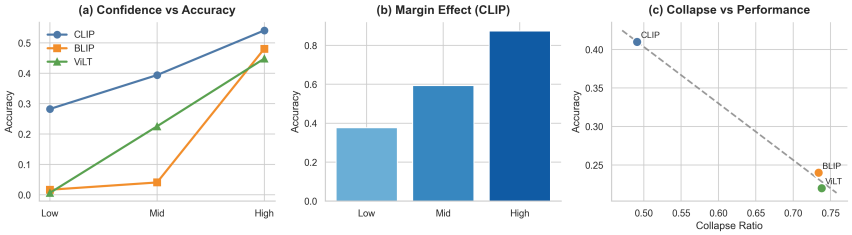


Figure 4: Prediction reliability under domain shift. (a) Confidence calibration. (b) Decision margins correlate better with correctness. (c) Accuracy decreases with collapse ratio.

To quantify this effect, we define the top-k concentration ratio:

$$R_k = \frac{\sum_{i=1}^{k} p_{(i)}}{\sum_{j=1}^{C} p_j},$$

where $p_{(i)}$ denotes sorted empirical frequencies over C candidate answers. For CLIP and ViLT, p is derived from softmax scores, while for BLIP it is estimated from normalized generation frequencies. A larger R_k indicates stronger output concentration. As shown in Figure 3(a)(b)(c), all models exhibit high concentration across domains. Moreover, predictions are dominated by a few frequent answers such as "yes", "no", and "abnormality", even for semantically diverse inputs. Figure 3(d) further reveals a many-to-one mapping from diverse ground-truth answers to a small set of prediction prototypes, indicating strong reliance on dataset-level priors under domain shift.

Confidence Miscalibration and Decision Instability. We further analyze prediction reliability under domain shift. As shown in Figure 4(a), confidence becomes increasingly miscalibrated as task difficulty increases. We define the decision margin as the gap between the top prediction and its closest competitor. Figure 4(b) shows that margin-based signals correlate more strongly with correctness than raw confidence, indicating degraded calibration of absolute probabilities. Figure 4(c) further shows a clear negative trend between output collapse and accuracy, suggesting that excessive concentration destabilizes decision boundaries.

Quantifying Output Distribution Collapse. To systematically measure this phenomenon, we introduce the Collapse Index. Given a dataset with N samples and C answer categories, we construct the empirical prediction distribution:

$$p = (p_1, p_2, \ldots, p_C),$$

where $p_c = \frac{N_c}{N}$ and N_c is the number of predictions assigned to category c.

We define entropy as:

$$H(p) = -\sum_{c=1}^{C} p_c \log(p_c + \epsilon),$$

where $\epsilon = 10^{-8}$. The entropy is bounded by:

$$0 \leq H(p) \leq \log C.$$

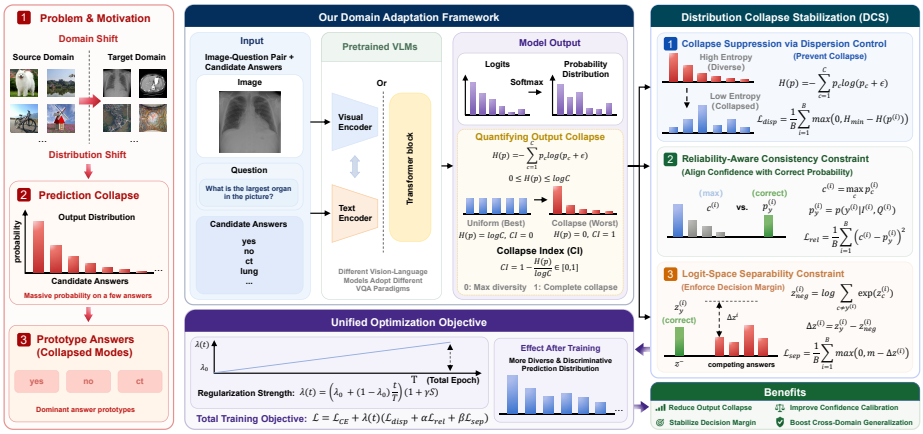


Figure 5: Overview of the proposed output-space regularization framework for cross-domain VQA.

The upper bound is achieved when:

$$p_c = \frac{1}{C}, \quad \forall c,$$

corresponding to maximal prediction diversity. The lower bound is attained when all predictions collapse into a single category. We define CI as:

$$CI = 1 - \frac{H(p)}{\log C} = \frac{\log C - H(p)}{\log C}.$$

Thus:

$$0 \leq CI \leq 1.$$

Lower CI indicates diverse predictions, while higher CI indicates stronger collapse. CI is computed over the full evaluation set and serves only as a diagnostic metric. We empirically find that higher CI correlates with lower cross-domain VQA performance.

Overall, as shown in Figures 3 and 4, domain shift induces a unified failure pattern: output distributions become concentrated, confidence becomes miscalibrated, and decision margins degrade, jointly reducing multimodal reasoning robustness. Motivated by this observation, we propose Distribution Collapse Stabilization, a unified output-space learning framework in the next section.

3 METHOD

We propose Distribution Collapse Stabilization (DCS), a unified output-space learning framework that jointly suppresses output distribution collapse, improves confidence calibration, and enhances logit-space separability under domain shift, as shown in Figure 5.

Prediction Distribution under Domain Shift. Given an image-question pair (I, Q), a vision-language model f_θ produces logits over the candidate answer set:

$$\mathbf{z} = f_\theta(I, Q) \in \mathbb{R}^C,$$

where C is the number of candidate answers. The predictive distribution is obtained via softmax:

$$p_c = \frac{\exp(z_c)}{\sum_{j=1}^{C} \exp(z_j)}.$$

We denote the ground-truth label as $y \in \{1, \ldots, C\}$. All losses are computed per sample and averaged over a batch.

Unified View of Output Distribution Collapse. We characterize a well-behaved predictive distribution through three properties: sufficient dispersion, calibration consistency, and logit-space separability. These correspond to different manifestations of a single requirement: maintaining a stable predictive distribution under domain shift.

Collapse Suppression via Dispersion Control. We measure distribution dispersion using entropy:

$$H(p) = -\sum_{c=1}^{C} p_c \log(p_c + \epsilon),$$

where $\epsilon = 10^{-8}$. To prevent excessive concentration, we enforce a hinge constraint:

$$\mathcal{L}_{disp} = \frac{1}{B}\sum_{i=1}^{B} \max(0, H_{\min} - H(p^{(i)})).$$

We set $H_{\min} = 0.3 \log C$ for all experiments, corresponding to a fixed fraction of the maximum entropy. This penalizes overly concentrated predictions while leaving diverse distributions unchanged.

Reliability-Aware Consistency Constraint. We define model confidence as:

$$c^{(i)} = \max_c p_c^{(i)},$$

and the ground-truth probability as:

$$p_y^{(i)} = p(y^{(i)} \mid I^{(i)}, Q^{(i)}).$$

To encourage calibrated confidence estimation, we minimize the discrepancy between the predicted confidence and the probability assigned to the correct answer:

$$\mathcal{L}_{rel} = \frac{1}{B}\sum_{i=1}^{B} \left(c^{(i)} - p_y^{(i)}\right)^2.$$

Logit-Space Separability Constraint. We define the ground-truth logit: $z_y^{(i)}$. The aggregated negative logit is computed using a log-sum-exp operator:

$$z_{neg}^{(i)} = \log \sum_{c \neq y^{(i)}} \exp(z_c^{(i)}).$$

The resulting decision margin is:

$$\Delta z^{(i)} = z_y^{(i)} - z_{neg}^{(i)}.$$

We enforce a minimum decision margin between the correct answer and all competing answers:

$$\mathcal{L}_{sep} = \frac{1}{B}\sum_{i=1}^{B} \max(0, m - \Delta z^{(i)}),$$

where $m = 0.5$. This mitigates decision-boundary instability under domain shift.

Collapse-Aware Scheduling Strategy. We define normalized entropy:

$$\bar{H}(p^{(i)}) = \frac{H(p^{(i)})}{\log C}, \quad s(p^{(i)}) = 1 - \bar{H}(p^{(i)}).$$

The batch-level collapse severity is defined as:

$$S = \frac{1}{B}\sum_{i=1}^{B} s(p^{(i)}).$$

The regularization weight is increased according to both training progress and the observed collapse severity:

$$\lambda(t) = \left(\lambda_0 + (1 - \lambda_0)\frac{t}{T}\right)(1 + \gamma S),$$

where $\lambda_0 = 0.2$, $\gamma = 0.4$, and t and T denote the current and total training steps, respectively.

Unified Objective. The overall DCS regularizer is defined as:

$$\mathcal{L}_{DCS} = \mathcal{L}_{disp} + \alpha\mathcal{L}_{rel} + \beta\mathcal{L}_{sep}.$$

The final training objective is:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda(t)\mathcal{L}_{DCS}.$$

We set $\alpha = 0.4$ and $\beta = 0.2$ for all experiments, while $\lambda(t)$ is updated dynamically during training.

4 EXPERIMENTS

4.1 DATASETS

We evaluate DCS on six representative cross-domain VQA benchmarks spanning medical imaging, remote sensing, and art understanding, including VQA-RAD Lau et al. (2018), SLAKE Liu et al. (2021b), PathVQA He et al. (2020), OVQA Huang et al. (2022), EarthVQA Wang et al. (2024), and SemArt Garcia et al. (2020).

4.2 BASELINES

(1) Domain-specific methods. For medical VQA, we compare representative methods, including MMQ Do et al. (2021), MEVF Nguyen et al. (2019), CR Zhan et al. (2020), CPRD Liu et al. (2021a), MMBert Khare et al. (2021), Med-CLIP Wang et al. (2022b), M3AE Chen et al. (2022), and PMC-CLIP Lin et al. (2023). For remote sensing and art understanding, we compare against SAN Yang et al. (2016), BAN Kim et al. (2018), BUTD Anderson et al. (2018), RSVQA Zheng et al. (2021), SOBA Wang et al. (2024), and VIKING Garcia et al. (2020). These methods leverage domain-specific architectures or domain-specific pretraining resources.

(2) General pretrained vision-language models. We further evaluate six representative pretrained vision-language models with substantially different architectures and training paradigms, including ViLT Kim et al. (2021), CLIP Radford et al. (2021), BLIP Li et al. (2022), Coca Yu et al. (2022), GIT Wang et al. (2022a), and Uform unum-cloud (2023). These models serve as generic VQA backbones and provide a challenging testbed for cross-domain adaptation.

(3) Adaptation baselines. To evaluate parameter-efficient adaptation under domain shift, we compare our DCS with FSA Liu et al. (2025), a representative feature-space adaptation framework. FSA introduces additional trainable modules into the backbone, whereas DCS directly regularizes the predictive distribution without introducing any additional parameters. FSA is selected as the primary adaptation baseline because it is one of the few methods designed for cross-domain VQA.

4.3 IMPLEMENTATION DETAILS

DCS is implemented as an output-space regularization module without modifying backbone architectures. All models are trained using AdamW, an initial learning rate of 2×10^{-5}. Experiments are conducted on NVIDIA V100 GPUs following the original training protocol of each backbone. Following prior cross-domain VQA studies, we adopt a unified candidate-answer evaluation protocol. The candidate pool is constructed from the training split using the most frequent answers, and the same candidate set is shared across all models to ensure fair comparison. For discriminative models, prediction scores are computed directly over the candidate set, while for generative models, we rank candidate answers using normalized conditional likelihoods.

4.4 MAIN RESULTS AND ANALYSIS

Tables 1 and 2 report the main results. DCS consistently improves all six backbone models across six cross-domain VQA benchmarks, achieving an average gain of 4.2% over the original pretrained models without introducing additional trainable parameters. Compared with FSA, DCS achieves comparable or superior performance despite requiring zero adaptation parameters. On medical benchmarks, DCS consistently improves performance on VQA-RAD, SLAKE, PathVQA,

Table 1: Cross-domain VQA performance on medical datasets. † denotes FSA-based adaptation and ∗ denotes the proposed Distribution Collapse Stabilization. DCS consistently improves all backbone models without introducing additional trainable parameters, yielding an average gain of 4.2% over the original pretrained baselines.

Methods	Radiology Domain						Pathology Domain			Cross Orthopedic		
	VQA-RAD			SLAKE			PathVQA			OVQA		
	Open	Closed	Overall	Open	Closed	Overall	Open	Closed	Overall	Open	Closed	Overall
MMQ	53.7	75.8	67.0	-	-	-	13.4	84.0	48.8	56.9	76.2	68.5
MEVF	49.2	77.2	66.1	77.8	79.8	78.6	8.1	81.4	44.8	36.3	76.3	60.4
CR	60.0	79.3	71.6	77.8	79.8	78.6	-	-	-	51.6	77.7	67.7
CPRD	52.5	77.9	67.8	79.5	83.4	81.1	-	-	-			
MMBert	63.1	77.9	72.0	-	-	-	-	-	-	37.9	80.2	63.3
Med-CLIP	60.1	80.0	72.1	78.4	82.5	80.1	-	-	-	-	-	-
M3AE	**67.2**	83.4	77.0	80.3	87.8	83.2	-	-	-	-	-	-
PMC-CLIP	67.0	**84.0**	**77.6**	81.9	88.0	84.3	-	-	-	52.6	82.3	70.5
ViLT (118.9M)	39.6	75.5	61.3	77.7	85.1	80.6	28.3	84.1	56.3	39.6	73.6	60.0
ViLT † (+1.0%)	49.1	75.5	65.1	78.5	87.2	81.9	33.7	85.2	59.5	42.7	77.0	63.3
ViLT ∗ (+0.0%)	50.3	76.4	66.0	80.6	86.5	82.9	34.1	86.8	60.5	46.0	81.3	67.2
CLIP (151.4M)	54.2	74.9	66.7	78.8	87.2	82.1	31.9	84.1	58.1	53.0	81.4	70.1
CLIP † (+0.9%)	59.8	78.9	71.4	81.9	88.3	84.4	36.5	85.0	60.8	62.8	83.4	75.2
CLIP ∗ (+0.0%)	62.2	79.8	72.8	81.9	**89.2**	84.7	35.3	87.1	61.2	61.8	85.5	76.1
BLIP (386.0M)	52.8	75.4	66.5	79.2	85.5	81.7	29.5	86.2	58.0	48.5	79.5	67.1
BLIP † (+0.4%)	58.1	77.0	69.6	79.8	88.3	83.2	33.8	86.5	60.2	52.3	81.0	69.5
BLIP ∗ (+0.0%)	58.7	78.0	70.3	81.6	86.3	83.5	35.1	87.3	61.3	56.5	82.8	72.3
Coca (253.7M)	32.2	71.0	55.7	79.4	83.0	80.8	25.5	85.0	55.4	52.3	80.4	69.2
Coca † (+0.9%)	45.3	70.5	60.5	80.2	86.6	82.7	30.7	87.2	59.0	55.2	82.3	71.5
Coca ∗ (+0.0%)	50.3	71.3	62.9	81.1	86.8	83.3	32.4	86.6	59.6	56.1	82.4	71.9
GIT (153.8M)	59.2	77.3	70.1	77.7	86.0	80.9	31.9	84.1	58.1	52.8	81.2	69.9
GIT † (+2.2%)	62.1	82.2	74.2	80.2	88.3	83.4	36.4	85.7	61.1	62.0	83.8	75.1
GIT ∗ (+0.0%)	65.4	82.8	76.0	**82.8**	88.3	**85.0**	34.9	87.7	61.4	**66.4**	**86.1**	**78.2**
Uform (1,500M)	55.8	76.2	68.2	78.7	83.6	80.6	32.0	86.4	59.3	51.3	79.9	68.5
Uform † (+0.2%)	59.2	79.6	71.6	81.0	85.4	82.7	**37.4**	85.7	61.6	58.1	81.9	72.4
Uform ∗ (+0.0%)	61.5	80.2	72.8	80.7	87.2	83.2	34.7	**89.8**	**62.4**	58.4	84.2	73.9

Table 2: Cross-domain VQA performance on remote sensing and art datasets, with FSA (†) and our DCS (∗).

Method	Remote	Art
	EarthVQA	SemArt
SAN	75.7	-
BAN	76.7	22.4
BUTD	76.4	21.8
Instruct-BLIP	75.2	-
RSVQA	70.8	-
SOBA	**78.1**	-
VIKING	-	20.4
ViLT	71.4	23.9
ViLT †	74.3	28.9
ViLT ∗	75.2	29.8
CLIP	74.2	26.4
CLIP †	76.3	29.6
CLIP ∗	76.8	30.3
BLIP	73.9	26.9
BLIP †	76.9	30.4
BLIP ∗	77.4	30.8
Coca	70.4	24.8
Coca †	73.7	28.9
Coca ∗	74.6	29.6
GIT	74.8	28.3
GIT †	77.6	31.1
GIT ∗	**78.1**	**31.4**
Uform	74.2	26.8
Uform †	76.8	29.1
Uform ∗	77.4	31.2

Table 3: Cross-domain VQA results (%) reported as mean ± standard deviation over three runs.

Methods	CLIP			BLIP			GIT		
	-	+FSA	+DCS	-	+FSA	+DCS	-	+FSA	+DCS
VQA-RAD	66.7±0.5	71.4±0.6	72.8±0.5	66.5±0.3	69.6±0.4	70.3±0.3	70.1±0.4	74.2±0.4	76.0±0.4
SLAKE	82.1±0.4	84.4±0.6	84.7±0.5	81.7±0.4	83.2±0.5	83.5±0.4	80.9±0.3	83.4±0.5	85.0±0.4
OVQA	70.1±0.3	75.2±0.6	76.1±0.6	67.1±0.3	69.5±0.4	72.3±0.3	69.9±0.3	75.1±0.6	78.2±0.4
SemArt	26.4±0.3	29.6±0.5	30.3±0.4	26.9±0.4	30.4±0.6	30.8±0.4	28.3±0.4	31.1±0.5	31.4±0.3

and OVQA. For example, CLIP improves from 66.7% to 72.8% on VQA-RAD and from 70.1% to 76.1% on OVQA. DCS also generalizes well beyond medical domains. On EarthVQA, GIT+DCS achieves 78.1%, matching the performance of the specialized SOBA model, while on SemArt it establishes a new state of the art with 31.4% accuracy. These results demonstrate the effectiveness and generality of output-space regularization under domain shift. To evaluate statistical robustness, Table 3 reports the mean accuracy and standard deviation. DCS consistently outperforms both the pretrained baselines and FSA across all datasets and architectures while exhibiting low variance, indicating that the observed improvements are stable and reproducible.

4.5 Extending DCS to Recent Large VLMs

Although DCS is motivated by the failure analysis of conventional VQA models, it regularizes predictive output distributions and is therefore architecture-agnostic. To evaluate its generality, we further apply DCS to four representative large VLMs, including LLaVA Liu et al. (2024b), BLIP-2 Li et al. (2023), Qwen3-VL Bai et al. (2025), and InternVL3.5 Wang et al. (2025). For autoregressive VLMs, we follow the same candidate-answer evaluation protocol by normalizing candidate likelihoods into predictive distributions, where likelihoods are computed by averaging token log-probabilities to avoid sequence-length bias. This enables DCS to be applied without modifying its optimization objective. Table 4 shows that DCS consistently improves all four VLMs across all datasets, achieving an average gain of approximately 4.0% over the corresponding pretrained baselines while outperforming FSA. The consistent improvements in recent autoregressive VLMs indicate that output distribution collapse is largely architecture-independent. Moreover, stronger backbones such as Qwen3-VL and InternVL3.5 also benefit from DCS, suggesting that increasing model capacity alone does not eliminate prediction collapse under domain shift.

4.6 Does DCS Improve Output Distribution Stability?

To verify that DCS improves predictive stability, we report Expected Calibration Error (ECE), Top-5 concentration ratio, and the collapse index (CI). As shown in Table 5, DCS consistently improves the reported stability metrics across datasets and backbones. Compared with the original models,

Table 4: Cross-domain VQA performance (%) on four representative large vision-language models.

Models	LLaVA			BLIP-2			Qwen3-VL			InternVL3.5		
	-	+FSA	+DCS	-	+FSA	+DCS	-	+FSA	+DCS	-	+FSA	+DCS
VQA-RAD	67.0	68.8	70.2	67.2	69.1	70.6	68.5	72.2	74.0	68.2	72.6	**74.4**
SLAKE	80.9	81.9	83.1	81.4	82.9	84.0	81.2	82.8	83.8	81.4	83.6	**84.4**
OVQA	68.6	70.7	73.0	69.2	72.8	75.0	70.6	73.2	74.9	70.4	74.1	**76.0**
SemArt	26.8	29.1	30.4	27.6	30.2	31.1	27.4	29.9	**31.2**	28.3	30.6	**31.2**

Table 5: Verification of output distribution stabilization. We report accuracy, Expected Calibration Error (ECE), Top-5 concentration ratio, and the normalized collapse index CI.

Dataset	Model	Acc. ↑	ECE ↓	Top-5 Ratio ↓	CI ↓
VQA-RAD	CLIP	66.7	0.302	72.3	0.58
	CLIP+DCS	72.8	0.168	58.6	0.44
	GIT	70.1	0.192	68.5	0.54
	GIT+DCS	**76.0**	**0.081**	**54.1**	**0.39**
OVQA	CLIP	70.1	0.156	64.8	0.52
	CLIP+DCS	76.1	0.081	52.2	0.42
	GIT	69.9	0.149	61.5	0.52
	GIT+DCS	**78.2**	**0.074**	**47.6**	**0.39**

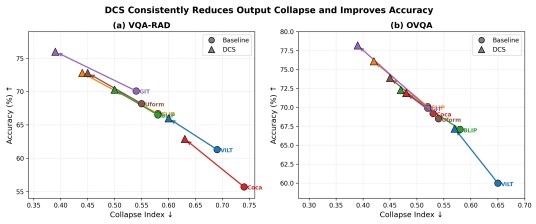


Figure 6: Relationship between collapse index and cross-domain VQA accuracy.

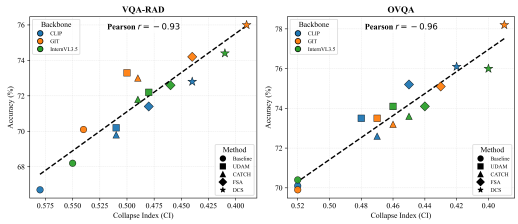


Figure 7: CI–accuracy correlation across adaptation methods and vision-language backbones.

DCS reduces prediction concentration, lowers CI, reduces calibration error, and simultaneously increases VQA accuracy. These results indicate that DCS effectively suppresses output distribution collapse while producing better-calibrated predictions. To further investigate this relationship, Figure 6 plots model accuracy against CI across different architectures on VQA-RAD and OVQA. A clear negative correlation can be observed: models with lower CI values consistently achieve higher accuracies. Moreover, DCS systematically shifts all backbones toward the upper-left region of the plot, corresponding to reduced collapse and improved performance. This trend is consistent across architectures and datasets, suggesting that the observed gains are closely associated with mitigating output distribution collapse rather than arising from generic regularization effects.

4.7 CI Correlates with Cross-Domain Accuracy

To examine whether this relationship generalizes across adaptation methods, we compare DCS against UDAM Weng et al. (2025), CATCH Li et al. (2025), and FSA Liu et al. (2025). Figure 7 shows a consistent negative correlation between CI and accuracy on both datasets. A clear negative correlation can be observed on both VQA-RAD and OVQA: models achieving lower CI consistently obtain higher accuracies. Although existing adaptation methods are primarily designed to improve feature representations, they also reduce CI compared with the original backbone models. This observation suggests that part of their performance gain may stem from indirectly alleviating output distribution collapse. In contrast, DCS achieves the lowest CI and highest accuracy across all settings, highlighting the advantage of directly regulating predictive output distributions.

4.8 Component-wise Analysis of DCS with CI

To further analyze the contribution of each DCS component from a distributional perspective, we introduce the CI as a diagnostic metric to quantify the severity of output distribution collapse. Unlike accuracy, which reflects task-level correctness, CI measures the degree of probability concentration in the output space. Importantly, CI is used only for post-hoc analysis and is not involved in opti-

Table 6: Ablation study of DCS on SLAKE and OVQA. We report accuracy (↑) and CI (↓).

$\mathcal{L}_{disp}$	$\mathcal{L}_{rel}$	$\mathcal{L}_{sep}$	$\lambda(t)$	CLIP		GIT	
				SLAKE (Acc / CI)	OVQA (Acc / CI)	SLAKE (Acc / CI)	OVQA (Acc / CI)
				82.1 / 0.32	70.1 / 0.52	80.9 / 0.35	69.9 / 0.52
✓				83.4 / 0.29	73.0 / 0.48	82.6 / 0.31	73.8 / 0.44
	✓			82.8 / 0.31	72.4 / 0.48	82.2 / 0.32	73.4 / 0.46
		✓		83.0 / 0.29	72.2 / 0.50	82.0 / 0.32	73.2 / 0.47
			✓	82.7 / 0.31	71.2 / 0.50	82.0 / 0.33	71.4 / 0.50
✓	✓			83.4 / 0.29	74.9 / 0.44	83.6 / 0.28	75.9 / 0.41
✓		✓		83.8 / 0.28	75.2 / 0.43	84.0 / 0.29	76.4 / 0.41
	✓	✓		83.2 / 0.29	73.6 / 0.46	83.2 / 0.29	75.2 / 0.43
✓	✓	✓		84.0 / 0.28	75.8 / **0.42**	84.4 / 0.28	77.4 / 0.41
✓	✓	✓	✓	**84.7 / 0.27**	**76.1 / 0.42**	**85.0 / 0.27**	**78.2 / 0.39**

Table 7: Comparison with standard regularization methods on CLIP. DCS consistently outperforms representative regularization strategies.

Dataset	CE	+Label Smoothing	+Focal Loss	+Temp. Scaling	+Entropy Reg.	+DCS
VQA-RAD	66.7	68.2	68.5	68.5	68.9	**72.8**
OVQA	70.1	72.0	72.3	72.3	72.6	**76.1**
SemArt	26.4	27.3	27.8	27.5	27.8	**30.3**

Table 8: Computational overhead of DCS on CLIP.

Method	Epoch Time (s)	Training Overhead	Inference Overhead
CE	58.9	–	–
Label Smoothing	59.1	+0.3%	0%
Focal Loss	59.1	+0.3%	0%
Entropy Reg.	59.2	+0.5%	0%
DCS	59.3	+0.6%	0%

mization. We conduct ablation studies on CLIP and GIT across SLAKE and OVQA. As shown in Table 6, we progressively enable dispersion control ($\mathcal{L}_{disp}$), reliability alignment ($\mathcal{L}_{rel}$), separability regularization ($\mathcal{L}_{sep}$), and the collapse-aware scheduling strategy $\lambda(t)$. Each component consistently reduces CI while improving accuracy, indicating that dispersion control, reliability alignment, and separability regularization contribute complementary benefits. Combining all components achieves the lowest CI and the highest accuracy, while the collapse-aware scheduling strategy further improves optimization stability. These results demonstrate that lower CI consistently corresponds to better cross-domain VQA performance.

4.9 Comparison with Regularization Methods

To investigate whether the gains of DCS can be attributed merely to generic regularization effects, we compare DCS with several widely used regularization strategies, including Label Smoothing Szegedy et al. (2016), Focal Loss Lin et al. (2017), Temperature Scaling Guo et al. (2017), and Entropy Regularization Pereyra et al. (2017). These methods represent four common directions for improving model robustness, including confidence smoothing, hard-sample reweighting, calibration correction, and prediction diversification. As shown in Table 7, all methods outperform the cross-entropy baseline, while DCS consistently achieves the best performance across all datasets. Compared with existing regularizers that mainly address a single aspect of prediction behavior (e.g., confidence smoothing, calibration, or entropy), DCS jointly regularizes predictive diversity, reliability, and decision separability, resulting in substantially larger performance gains under domain shift. Table 8 reports the computational overhead of DCS. Since all regularization terms are computed directly from the predictive logits, DCS introduces no additional parameters and incurs only 0.6% extra training time with zero inference overhead, making it a lightweight plug-and-play strategy.

5 Conclusion

In this paper, we revisit cross-domain visual question answering from the perspective of output-space behavior. Through systematic analyses across multiple domains and vision-language architectures, we identify a unified failure mechanism, termed output distribution collapse, characterized by concentrated predictive distributions, confidence miscalibration, and unstable decision margins. Motivated by this observation, we propose DCS, a lightweight output-space regularization framework that jointly preserves distribution diversity, confidence reliability, and decision separability without modifying model architectures or introducing additional trainable parameters. Extensive experiments on six cross-domain VQA benchmarks demonstrate consistent improvements across diverse vision-language backbones, yielding an average accuracy gain of 4.1% with negligible computational overhead. We hope this work highlights the importance of output-space modeling for cross-domain VQA and encourages future research on prediction-level adaptation in vision-language models.

AI USE STATEMENT

Generative AI tools were used for language editing and polishing, LaTeX formatting, table and figure layout adjustments, and limited assistance in editing plotting and formatting code. All AI-assisted outputs were reviewed and verified by the authors. Generative AI tools were not used to generate research ideas, develop the proposed method, design or conduct experiments, analyze experimental results, or formulate scientific claims. The authors take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

The paper and Appendix provide the method formulation, datasets, experimental settings, hyperparameters, implementation details, and additional experiments needed to reproduce the reported evaluation protocol and results.

REFERENCES

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

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Sergio A Balanya, Juan Maronas, and Daniel Ramos. Adaptive temperature scaling for robust calibration of deep neural networks. *Neural Computing and Applications*, 36(14):8073–8095, 2024.

Tinghui Chen, Weiyi Yang, Shuai Li, and Xin Luo. Data-driven calibration of industrial robots: a comprehensive survey. *IEEE/CAA Journal of Automatica Sinica*, 12(8):1544–1567, 2025.

Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*, pp. 679–689. Springer, 2022.

Tuong Do, Binh X Nguyen, Erman Tjiputra, Minh Tran, Quang D Tran, and Anh Nguyen. Multiple meta-model quantifying for medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pp. 64–74. Springer, 2021.

Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906*, 2021.

Ziqi Gao, Qichao Wang, Aochuan Chen, Zijing Liu, Bingzhe Wu, Liang Chen, and Jia Li. Parameter-efficient fine-tuning with discrete fourier transform. *arXiv preprint arXiv:2405.03003*, 2024.

Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. A dataset and baselines for visual question answering on art. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 92–108. Springer, 2020.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Xuefeng Hu, Ke Zhang, Lu Xia, Albert Chen, Jiajia Luo, Yuyin Sun, Ken Wang, Nan Qiao, Xiao Zeng, Min Sun, et al. Reclip: Refine contrastive language image pre-training with source free domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2994–3003, 2024.

Yefan Huang, Xiaoli Wang, Feiyan Liu, and Guofeng Huang. Ovqa: A clinically generated visual question answering dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2924–2938, 2022.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pp. 709–727. Springer, 2022.

Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1033–1036. IEEE, 2021.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19113–19122, 2023a.

Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15190–15200, 2023b.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pp. 5583–5594. PMLR, 2021.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

Will LeVine, Benjamin Pikus, Pranav Raja, and Fernando Amat Gil. Enabling calibration in the zero-shot inference of large vision-language models. *arXiv preprint arXiv:2303.12748*, 2023.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Xinjin Li, Yulie Lu, Jinghan Cao, Yu Ma, Zhenglin Li, and Yeyang Zhou. Catch: A modular cross-domain adaptive template with hook. *arXiv preprint arXiv:2510.26582*, 2025.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 525–536. Springer, 2023.

Bo Liu, Li-Ming Zhan, and Xiao-Ming Wu. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 210–220. Springer, 2021a.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1650–1654. IEEE, 2021b.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.

Lei Liu, Xiangdong Su, and Guanglai Gao. Fourier self-adaptation for transferring general pre-trained models to specific domains. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3605–3614, 2025.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024c.

Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. *arXiv preprint arXiv:1612.02295*, 2016.

Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566, 2020.

Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. Overcoming data limitation in medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 522–530. Springer, 2019.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Deepak Ravikumar, Efstathia Soufleri, and Kaushik Roy. Improved out-of-distribution detection with additive angular margin loss. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3425–3432, 2025.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

unum-cloud. Pocket-sized multimodal ai for content understanding and generation. `https://github.com/unum-cloud/uform`, 2023.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022a.

Junjue Wang, Zhuo Zheng, Zihang Chen, Ailong Ma, and Yanfei Zhong. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5481–5489, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, pp. 3876, 2022b.

Weixi Weng, Rui Zhang, Xiaojun Meng, Jieming Zhu, Qun Liu, and Chun Yuan. Unsupervised domain adaptive visual question answering in the era of multi-modal large language models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6248–6258. IEEE, 2025.

Yiming Xu, Lin Chen, Zhongwei Cheng, Lixin Duan, and Jiebo Luo. Open-ended visual question answering by multi-modal domain adaptation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 367–376, 2020.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 21–29, 2016.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2345–2354, 2020.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

Xiangtao Zheng, Binqiang Wang, Xingqian Du, and Xiaoqiang Lu. Mutual attention inception network for remote sensing visual question answering. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022.