Mind the Gap: Preserving and Compensating for the Modality Gap in CLIP-Based Continual Learning

Supplementary Material

1. Visualization of Modality Gap

As shown in the Fig. 1, we randomly sampled 512 image-caption pairs from LAION-400M [3], extracted features using CLIP, and applied UMAP [2] for dimensionality reduction. The results reveal a clear clustering of features within the same modality, while features from different modalities maintain a certain distance. This phenomenon reflects the modality gap.

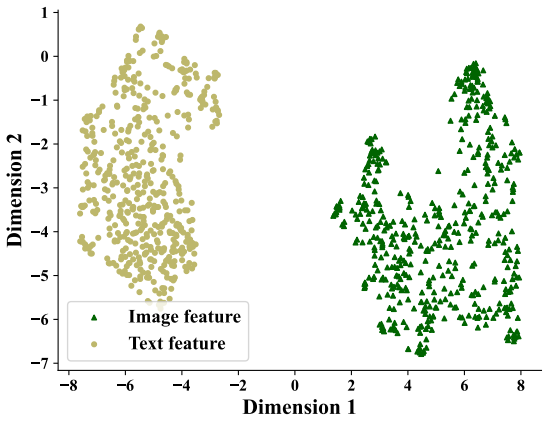

Figure 1. The features of the image and its corresponding caption visualized using UMAP.

2. Proof of Image-Space Classifier and Modality Gap Constraints

2.1. Existence of an Optimal Classifier within Image Feature Space

Let the CLIP image feature matrix be $\mathbf{X} \in \mathbb{R}^{d\times n}$. Consider a classifier $\mathbf{W}_* \in \mathbb{R}^{d\times C}$ that achieves minimal cross-entropy loss under ideal conditions. We show that there always exists an equivalent classifier $\mathbf{W}_{\text{opt}}$ that is entirely contained within the span of $\mathbf{X}$, i.e., $\mathbf{W}_{\text{opt}} \in \text{span}(\mathbf{X})$.

Since any classifier $\mathbf{W}_*$ can be decomposed as:

$$\mathbf{W}_* = \mathbf{W}_{\parallel} + \mathbf{W}_{\perp}, \tag{1}$$

where $\mathbf{W}_{\parallel} \in \text{span}(\mathbf{X})$ and $\mathbf{W}_{\perp} \perp \text{span}(\mathbf{X})$. The input feature $\mathbf{x}_i \in \mathbf{X}$ and its label y contribute to the loss function composed of softmax and cross-entropy as follows:

$$\mathcal{L}_{ce} = -\sum_{i}^{n} \log \frac{\exp(\mathbf{w}_y^\top \mathbf{x}_i)}{\sum_j^C \exp(\mathbf{w}_j^\top \mathbf{x}_i)} \tag{2}$$

Obviously, $\mathbf{W}_{\perp}$ does not relate to the loss. Thus, an equivalent classifier achieving the same loss can be expressed as:

$$\mathbf{W}_{\text{opt}}^\top \mathbf{X} = \mathbf{W}_{\parallel}^\top \mathbf{X} = \mathbf{W}_*^\top \mathbf{X}. \tag{3}$$

This establishes the existence of an optimal classifier contained within the image feature space.

2.2. Restriction Imposed by Modality Gap on Text Classifiers

Using Singular Value Decomposition (SVD), we express $\mathbf{W}_{\text{opt}}$ as:

$$\mathbf{W}_{\text{opt}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \tag{4}$$

where $\mathbf{U} \in \mathbb{R}^{d\times r'}$ represents an orthonormal basis for a subspace of span$(\mathbf{X})$. The text feature matrix $\mathbf{T}$ can be decomposed into:

$$\mathbf{T} = \mathbf{T}_{\parallel} + \mathbf{T}_{\perp}, \tag{5}$$

where $\mathbf{T}_{\parallel} = \mathbf{U}_r\mathbf{A}$ (with rank $r \le r'$) lies within the image feature subspace, and $\mathbf{T}_{\perp}$ is its orthogonal complement.

Since $\mathbf{T}_{\perp}$ does not contribute to classification, the optimal alignment is obtained by solving:

$$\min_{\mathbf{A}} \|\mathbf{U}_r\mathbf{A} - \mathbf{W}_{\text{opt}}\|_F^2. \tag{6}$$

The optimal solution is given by:

$$\mathbf{A}^* = \mathbf{U}_r^\top \mathbf{W}_{\text{opt}}. \tag{7}$$

Substituting this back, the misalignment error is lower-bounded by:

$$\|\mathbf{T}_{\parallel} - \mathbf{W}_{\text{opt}}\|_F^2 \ge \|\mathbf{U}_r\mathbf{U}_r^\top\mathbf{W}_{\text{opt}} - \mathbf{W}_{\text{opt}}\|_F^2 = \sum_{i=r+1}^{r'} s_i^2, \tag{8}$$

where s_i are the singular values of $\mathbf{W}_{\text{opt}}$.

This result implies that perfect alignment is achievable only when the text feature subspace has sufficient rank to fully capture $\mathbf{W}_{\text{opt}}$, i.e., when $r = r'$. However, due to the modality gap, the effective rank of the text classifier subspace is often lower ($r < r'$), leading to an inherent limitation in classification performance.

3. Implementation Details of Image Space and Classifier Space Analyzing

To analyze these relationships, we first apply SVD to the image feature matrix and extract its corresponding basis vectors, denoted as $\mathbf{B}_i \in \mathbb{R}^{d\times r}$. Given the large number of

	CIFAR100		ImageNet100	
	Avg	Last	Avg	Last
ours w/o replay	86.79	80.40	87.31	78.38
ours w/ replay	88.48	82.58	88.50	80.74

Table 1. Experimental results with replay data on our method

Method	CIFAR100		ImageNet-R	
	Avg	Last	Avg	Last
PROOF	89.87	83.59	91.25	87.33
CLAP	87.94	84.86	92.12	88.63
SLCA	90.12	84.62	89.99	86.83
RAPF	90.25	85.29	91.96	88.32
L2P++	85.68	77.86	90.49	86.73
DualPrompt	86.63	79.12	90.66	87.14
CODA	85.82	78.67	89.11	84.56
Continual-CLIP	80.48	73.46	86.99	83.05
Aper-Adapter	80.21	71.95	89.17	85.4
MOE4CL	90.98	85.83	93.27	90.42
CLAP*	74.41	71.57	91.10	87.55
MagMax	90.16	86.06	93.22	89.55
ours	91.78	87.03	93.66	91.08

Table 2. Experimental Results with ViT-L/14 Backbone Model

image features, we retain only the basis vectors that capture 95% of the total energy, reducing noise while preserving essential information. For the text feature classifier and the visual space classifier, we employ QR decomposition to obtain their respective basis vectors, denoted as $\mathbf{B}_t$ and $\mathbf{B}_{vc}$. Furthermore, we compute the basis vectors of the combined space spanned by both classifiers, represented as $\mathbf{B}_{t+vc}$.

4. More experiments

4.1. Compatibility with replay methods.

Our method does not require rehearsal samples; however, it is still compatible with them. We tested our method with simple random sampling of rehearsal data, keeping the total number at 2000. The experimental results, shown in Table 1, demonstrate that our method benefits from the replay data and is compatible with it.

4.2. Experiments of CLIP ViT-L/14 backbone

We evaluate the effectiveness of our method on another CLIP model. For this, we replace the backbone of all methods with OpenAI's stronger ViT-L/14 model. The experimental results, shown in Table 2, demonstrate that our method still outperforms the others.

Rank	4	8	16	32
Last	78.02	78.38	78.32	78.26

Table 3. Experiments with different ranks of LoRA on ImageNet-100

β	1	2	4	6	8
Last	77.58	77.92	78.38	77.98	77.32

Table 4. Experiments with different β on ImageNet-100

α	5%	10%	20%	30%
Last	77.88	78.38	77.34	76.9

Table 5. Experiments with different α on ImageNet-100

4.3. Hyperparameter selection

We use the same hyperparameters for all datasets. Below are the experiments for hyperparameter selection on a single dataset.

Rank of LoRA As shown in Table 3, our method is insensitive to the rank of LoRA. We choose a rank of 8 for experiments across all datasets.

Output the ensemble weights β As shown in Table 4, as we increase the weight β assigned to the visual-space classifier's output, the overall performance first improves and then declines. This suggests that higher-confidence predictions from the visual-space classifier effectively compensate for the shortcomings of the text classifier, while lower-confidence predictions have minimal impact on the overall results. However, when β becomes too large, even low-confidence predictions from the visual classifier can significantly influence the text classifier's output, allowing incorrect predictions with low scores to dominate. Therefore, an appropriately chosen weight enables better complementarity between the two classifiers.

Effect of α As shown in the Tab. 5, when α is too small, it excessively constrains model training, preventing it from fully learning the new task and limiting performance. Conversely, when α is too large, the pre-trained knowledge is disrupted, leading to a gradual performance decline. We select 10% as our α.

4.4. Zero-shot Capability

As shown in Tab. 6, we run zero-shot tests on four ImageNet-C [1] corruptions (severity 3) after continual

	Defocus	Contrast	Frost	Gaussian
CLIP	41.95	55.07	38.23	43.20
MagMax	43.51	52.10	36.75	41.43
MOE4CL	44.24	54.82	34.75	36.09
Ours	44.65	56.02	37.71	42.47

Table 6. Zero-shot performance on ImageNet-C after class-incremental learning on CIFAR-100.

	100-shot	50-shot	25-shot	5-shot
MagMax	75.82	75.02	72.53	67.66
MOE4CL	75.52	75.40	74.98	68.54
Ours	78.30	78.04	77.04	75.10

Table 7. Last accuracy on CIFAR-100 (10-task) in few-shot settings, showing consistent gains of our method under limited data.

	CIFAR-100		ImageNet-R	
	5task	20task	5task	20task
MagMax	82.07	76.84	82.75	80.18
MOE4CL	78.96	76.20	81.37	79.58
LGVLM[4]	83.84	77.26	82.46	79.32
Ours	81.47	79.31	83.13	82.12

Table 8. Last accuracy under different task settings on CIFAR-100 and ImageNet-R.

learning from CIFAR-100. **Defocus blur**: Due to CIFAR-100's low resolution, all methods slightly improved over the original CLIP; our method performed best. **Other corruptions**: Other methods showed reduced robustness, while ours maintained CLIP performance with minimal drop and slight gains on Contrast. This suggests our approach maintains CLIP's generalization better than other methods.

4.5. Few-shot Capability

As shown in Tab. 7, we report last accuracy in continual learning on CIFAR-100 10 task under limited data settings. Our method shows greater advantage under limited data conditions.

4.6. Additional task settings

We further evaluate our method under different task settings. As shown in Tab. 8, while our performance on CIFAR-100 (5-task) is lower than the best baseline, our method achieves the highest accuracy under the more challenging 20-task setting.

	Ours	+ EMA	+ Epoch Est.
Avg	87.58	87.73	87.77
Last	82.67	82.92	82.82

Table 9. Study of auxiliary strategies on ImageNet-R (10-task).

4.7. Study on Auxiliary Strategies

To explore potential performance enhancements, we test two auxiliary strategies: (1) **Exponential Moving Average (EMA)** and (2) **Per-task Epoch Estimation**. As shown in Tab. 9, both bring minor improvements but will increase computational cost. In particular, per-task epoch estimation doubles the number of forward passes. To maintain efficiency, we adopt our method without these additions.

5. Algorithm Pseudocode

The overall pipeline of training is shown in the pseudocode:

Algorithm 1 Training algorithm

Input: $\mathbf{D} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_t\}$ ▷ Training data in all tasks
Input: $f^0_{\text{clip}}(\cdot)$ ▷ Original CLIP
require: $f^T_{clip}(\cdot)$ ▷ CLIP after fine-tuning on T tasks
require: $\mathbf{W}_v$ ▷ cosine classifier in visual space
Initialize: $e = 0$ ▷ fine-tuning period
for $t = 1$ to T **do**
  $f^{t,0}_{\text{clip}}(\cdot) = f^{t-1}_{\text{clip}}(\cdot)$ ▷ init model in current task
  if $t = 1$ **then** ▷ calculate the fine-tuning period
    $neg^0 = Eq.2(f^{1,0}_{\text{clip}}(\cdot), \mathbf{X}_1)$
    $f^{1,1}_{\text{clip}}(\cdot) = FINETUNE(f^{1,0}_{\text{clip}}(\cdot), \mathbf{X}_1)$
    $neg^1 = Eq.2(f^{1,1}_{\text{clip}}(\cdot), \mathbf{X_1})$
    while $Eq.3(neg^0, neg^{e+1}) < \alpha$ **do**
      $e+ = 1$
      $f^{1,e+1}_{\text{clip}}(\cdot) = FINETUNE(f^{1,e}_{\text{clip}}(\cdot), \mathbf{X}_1)$
      $neg^{e+1} = Eq.2(f^{1,e+1}_{\text{clip}}, \mathbf{X_1})$
    $e = max(1, e)$
  else
    for $i = 1$ to e **do**
      $f^{t,i}_{\text{clip}}(\cdot) = FINETUNE(f^{t,i-1}_{\text{clip}}(\cdot), \mathbf{X}_t)$
  $f^t_{\text{clip}}(\cdot) = f^{t,e}_{\text{clip}}(\cdot)$
  Initialize: $\mathbf{W}^t_v$ ▷ initalize $\mathbf{W}^t_v$ based on class prototypes
  $W^t_v = TRAIN(\mathbf{W}^t_v, f^t_{\text{clip}}(\mathbf{X_t}))$
return $f^T_{\text{clip}}(\cdot), \mathbf{W}^T_v$

References

[1] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. 2

[2] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 1

[3] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1

[4] Wentao Zhang, Yujun Huang, Weizhuo Zhang, Tong Zhang, Qicheng Lao, Yue Yu, Wei-Shi Zheng, and Ruixuan Wang. Continual learning of image classes with language guidance from a vision-language model. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 3