

Disentangling CLIP Features for Enhanced Localized Understanding

Samyak Rawlekar¹ Yujun Cai² Yiwei Wang³ Ming-Hsuan Yang^{3,4} Narendra Ahuja¹

Abstract

Vision-language models (VLMs) demonstrate impressive capabilities in coarse-grained tasks like image classification and retrieval. However, they struggle with fine-grained tasks that require localized understanding. To investigate this weakness, we comprehensively analyze CLIP features and identify an important issue: semantic features are highly correlated. Specifically, the features of a class encode information about other classes, which we call mutual feature information (MFI). This mutual information becomes evident when we query a specific class and unrelated objects are activated along with the target class. To address this issue, we propose Unmix-CLIP, a novel framework designed to reduce MFI and improve feature disentanglement. We introduce MFI loss, which explicitly separates text features by projecting them into a space where inter-class similarity is minimized. To ensure a corresponding separation in image features, we use multi-label recognition (MLR) to align the image features with the separated text features. This ensures that both image and text features are disentangled and aligned across modalities, improving feature separation for downstream tasks. For the COCO-14 dataset, Unmix-CLIP reduces feature similarity by 24.9%. We demonstrate its effectiveness through extensive evaluations of MLR and zero-shot semantic segmentation (ZS3). In MLR, our method performs competitively on the VOC2007 and surpasses SOTA approaches on the COCO-14 dataset, using fewer training parameters. Additionally, Unmix-CLIP consistently outperforms existing ZS3 methods on COCO and VOC.

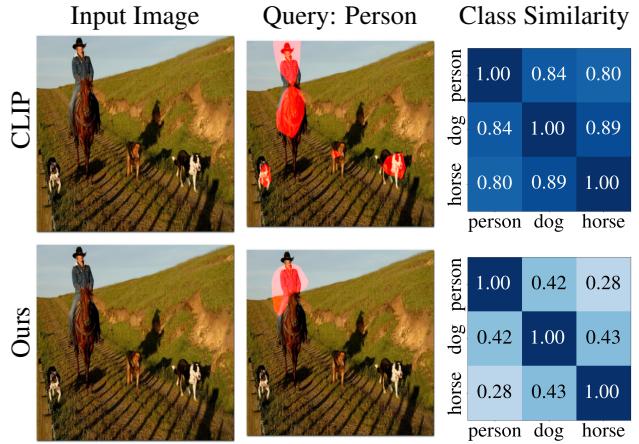


Figure 1: Comparison of Activated Regions. When queried for the ‘person’ class (middle column, highlighted in red), CLIP shows activation in unqueried regions (dogs and horses), while our method maintains focus on the person. The rightmost column displays cosine similarities between class features, showing that reducing the inter-class similarity (person-dog: $0.84 \rightarrow 0.42$, person-horse: $0.80 \rightarrow 0.28$) results in features that are suitable for fine-grained tasks.

1. Introduction

Vision-language models (VLMs) have emerged as powerful tools for understanding visual content through natural language supervision. CLIP (Radford et al., 2021), trained on 400 million image-text pairs (WIT-400M), achieves remarkable performance in coarse-grained visual understanding tasks such as image classification (Zhou et al., 2022b), image retrieval (Baldrati et al., 2022), and visual question answering (Yu et al., 2022). However, these models struggle with fine-grained tasks that require localized understanding, leading to significant performance degradation in multi-label recognition (MLR) (Zhu & Wu, 2021; Huang et al., 2024) and semantic segmentation (Lüdecke & Ecker,

¹Department of Electrical and Computer Engineering, University of Illinois Urbana-Champaign, USA ² University of Queensland, Brisbane, Australia ³UC Merced, USA ⁴Yonsei University, South Korea. Correspondence to: Samyak Rawlekar <samyakr2@illinois.edu>.

2022). While previous work has attributed these limitations to architectural choices (Dariset et al., 2023; Zhou et al., 2022a; Li et al., 2023) or training objectives (Lin et al., 2024; Dong et al., 2023), our analysis reveals a more fundamental issue: the entanglement of semantic features in CLIP’s feature space.

We systematically analyze CLIP’s features and identify two key factors contributing to this issue. First, the spatial pooling operation in the final layer, although effective for global tasks, discards essential localized information necessary for fine-grained understanding. Second, and more importantly, we discover significant interference between class features in the joint vision-language space, which we term mutual feature information (MFI). The mutual information becomes apparent during class-specific queries: regions corresponding to unrelated objects are consistently activated alongside the target class. For example, as illustrated in Figure 1, regions containing ‘dog’ and ‘horse’ also activate when we query the class ‘person.’ This activation pattern strongly correlates with the high similarity scores between class text features (0.84 for person-dog and 0.80 for person-horse), indicating substantial feature entanglement in CLIP’s representation space.

To address this fundamental limitation, we introduce Unmix-CLIP, a novel framework that disentangles class features in vision-language models. Drawing inspiration from the redundancy reduction principle (Barlow et al., 1961) in neuroscience, we extend this concept to the vision-language domain. While previous approaches have focused on architectural modifications (Zhou et al., 2022a; Li et al., 2023; Boushelham et al., 2024) or prompt engineering (Sun et al., 2022; Rawlekar et al., 2024a) to adapt VLMs for fine-grained tasks, Unmix-CLIP directly targets the root cause by minimizing MFI between class representations while preserving task-relevant information. We achieve this through a carefully designed MFI loss function that explicitly disentangles text features by projecting them to minimize inter-class similarity. To achieve a similar separation in image features, we align them with the projected text features using a multi-label recognition framework. The joint training using MFI loss (separates text features) and MLR loss (aligns text and image features) results in disentangled features that align across the image and text domains, leading to improved separation in semantic features.

We train Unmix-CLIP on 80 classes from the COCO-14 (Lin et al., 2014) dataset and evaluate its performance on two fine-grained tasks: multi-label recognition (MLR) and zero-shot semantic segmentation (ZS3). For MLR evaluation, we use the COCO-14 and VOC2007 (Everingham et al., 2010) datasets. For ZS3, we use VOC2012 (Everingham et al., 2010) and COCO-17 (Lin et al., 2014) for seen classes, and VOC Context (Mottaghi et al., 2014) provides

59 classes, 30 of which are unseen during pre-training. Our experimental results demonstrate that Unmix-CLIP achieves competitive performance on VOC and outperforms state-of-the-art (SOTA) methods on the challenging COCO-14 dataset, using only one-third of their training parameters. For ZS3, Unmix-CLIP surpasses SOTA VLMs on datasets with seen classes, demonstrating that reducing mutual feature information (MFI) is crucial for fine-grained tasks. To further assess its segmentation capabilities, we apply Unmix-CLIP to segment objects in the images, recasting the task as single-label recognition, a task more suitable for CLIP. We combine the segment-level and whole-image results to obtain zero-shot MLR predictions. Segmenting objects provides complementary information on top of global image features. The main **contributions** of this work are:

- We identify a critical challenge in adapting VLMs for fine-grained tasks: mutual information between class features (MFI) degrades fine-grained task performance
- To address this challenge, we propose Unmix-CLIP, a framework that adapts CLIP features for fine-grained tasks by reducing MFI. At its core lies our proposed MFI loss, which explicitly disentangles text features and guides the disentanglement of image features
- We show that Unmix-CLIP outperforms SOTA multi-label recognition methods in challenging settings using significantly fewer training parameters. Additionally, it outperforms zero-shot semantic segmentation methods. Moreover, as an object segmenter, Unmix-CLIP enhances CLIP’s zero-shot MLR performance.

2. Related Work

Recoding information. Shannon proposed that optimal information transmission involves designing codes with minimum entropy (Shannon, 1948). The redundancy reduction principle extended this idea to neuroscience, suggesting that sensory systems recode information to reduce redundancy with minimal loss (Barlow et al., 1961). This principle has since been applied to many recent works, including image compression (Ballé et al., 2016) and more popularly in representation learning (Oord et al., 2018; Chen et al., 2020; Zbontar et al., 2021; Henaff, 2020; He et al., 2020; Chen & He, 2021). While our loss function shares structural similarities with representation learning methods (a similarity and contrastive term), our method differs as follows: (1) Instead of learning features from scratch, we refine learned features (reducing MFI). (2) We do not rely on augmentation-based learning or batch processing. (3) Unlike contrastive methods that require paired embeddings, our approach operates on a fixed set of text embeddings. Most importantly, our objective is not generic feature learning but targeted feature modification to enhance task-specific utility.

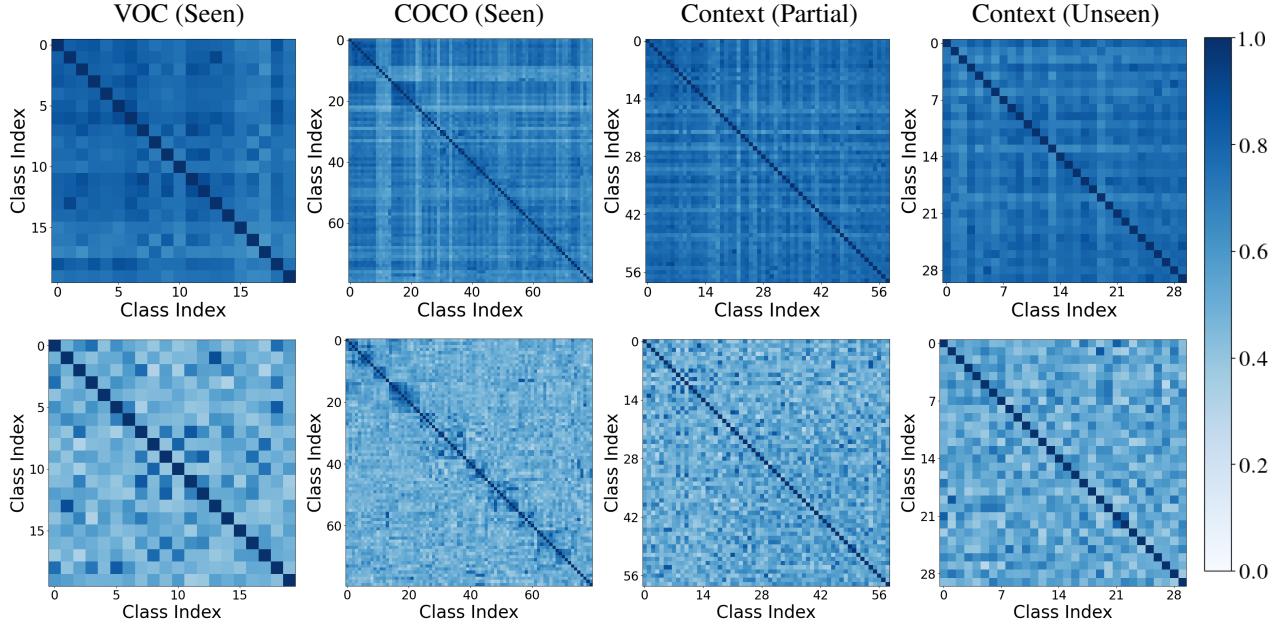


Figure 2: Class Feature Similarity Analysis. Comparison of class-text feature similarities between CLIP (top row) and our method (bottom row) across four datasets: VOC, COCO, Context (partial), and Context (unseen). The heatmaps show cosine similarity between class text features, where darker blue indicates higher similarity. As the reduced off-diagonal similarity values show, our method achieves higher class feature separation. This improved class separation suggests better discrimination capabilities.

Vision-Language Models for Fine-grained Tasks. Vision-language models (VLMs) trained with contrastive losses (Radford et al., 2021; Jia et al., 2021) are challenging to adapt for fine-grained tasks due to two reasons: (1) their reliance on global feature aggregation, which ignores local information. (2) Using the softmax operation in their training loss biases them toward single-object settings.

Recognition. Early efforts to adapt VLMs for recognition centered on learning prompts as classifiers for visual features (Zhou et al., 2022b). These methods were extended to multi-label settings by learning multiple prompts for each class (Sun et al., 2022; Hu et al., 2023; Rawlekar et al., 2024a). Subsequent works incorporated co-occurrence information to make predictions interdependent (Ding et al., 2023; Rawlekar et al., 2024b). In contrast, our approach does not rely on prompt learning or co-occurrence modeling during pre-training. Furthermore, our features are adaptable to tasks beyond multi-label recognition.

Localization. Early approaches addressed localization by training image segmentation models and using VLMs to label the segmented regions (Kirillov et al., 2023). Later methods introduced pre-training setups that combined vision-language alignment with mask distillation to enhance localization (Dong et al., 2023). Recent works adapted features for localization without additional training by leveraging

the spatial properties preserved in the value projection of CLIP’s transformer-style aggregation (Zhou et al., 2022a). CLIP Surgery (Li et al., 2023) identified consistent noisy activations across classes and reduced them by subtracting average features from class-specific features (Li et al., 2023), though the cause of these activations remains unclear. GEM generalized this concept to vision transformers (Bousselham et al., 2024). We use the finding that value projection preserves spatial information. We further improve value projection by disentangling class features.

3. Unmix-CLIP

Given a multi-label dataset \mathcal{D} , where $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}|}$ consists of images \mathbf{x}_i and N class labels $\{C_j\}_{j=1}^N$, each image \mathbf{x}_i can contain objects belonging to one or more of these N classes. Additionally, we use CLIP (f_θ), parameterized by weights θ , consisting of an image encoder ($f_{\theta,\text{img}}$) and a text encoder ($f_{\theta,\text{text}}$) for feature extraction. Throughout all experiments, we keep the parameters of CLIP (f_θ) frozen, including both the image and text encoders.

Since the mutual information among class features present in CLIP is detrimental to fine-grained task performance, we analyze CLIP’s features from this perspective. Specifically, we focus on two key aspects: (1) spatial preservation in the visual feature maps and (2) the relationship between class

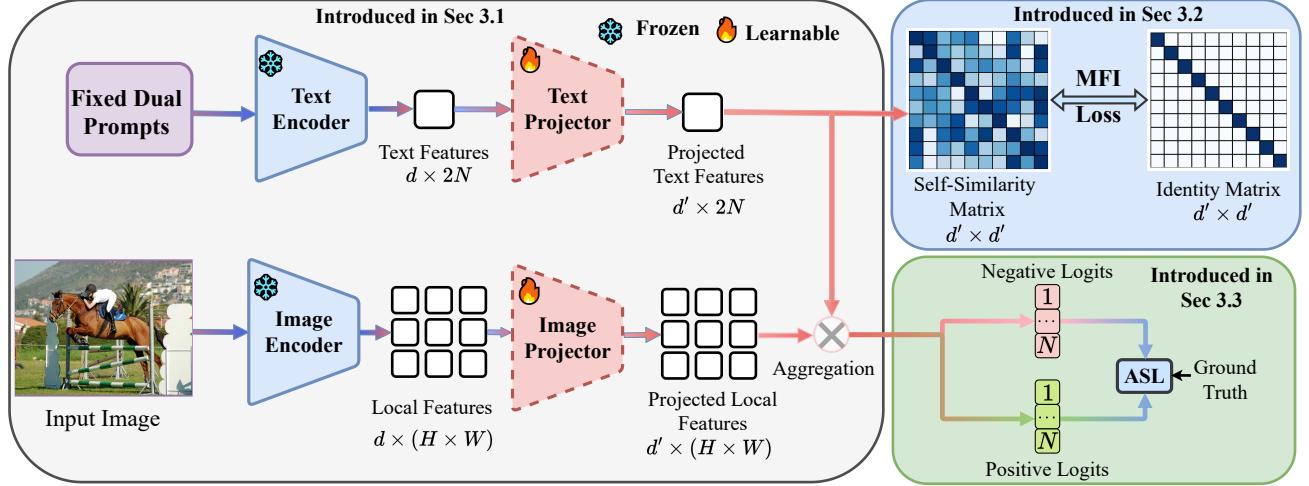


Figure 3: Unmix-CLIP Overview. Given image and label names in the dataset, CLIP extracts image and text features, which are further processed by respective projectors to embed into a disentangled space while preserving local image information. To reduce mutual feature information (MFI) between class features, we propose MFI loss that enforces the self-similarity matrix of projected text features to approximate an identity matrix, effectively reducing inter-class feature dependencies (Section 3.2). We propagate the separation in the text features to image space by aligning the image and separated text features using a multi-label recognition setup (Section 3.3). Following (Sun et al., 2022) and as detailed in Section 3.3, we aggregate the projected image and text features to obtain predicted logits. The predicted logits are trained with ground truth labels using the widely used asymmetric loss (ASL) (Ridnik et al., 2021). Our training loss combines the ASL and MFI loss; the only trainable components are the projectors. We freeze both CLIP encoders and projectors during inference for multi-label recognition and downstream tasks such as zero-shot semantic segmentation.

features in the joint vision-language space.

Towards (1), we remove CLIP’s final spatial pooling layer to preserve local information in feature maps. We then evaluate class-wise activations by computing the similarity between local visual and text features. For (2), we find that querying an image for a specific class consistently activates unrelated regions. Figure 1 shows that querying for ‘person’ highlights the person regions and activates areas containing dogs and horses. This suggests that CLIP’s features for different classes share substantial information.

To quantify this feature entanglement, we analyze the similarity between class text features across multiple datasets (VOC (Everingham et al., 2010), COCO (Lin et al., 2014), and Context (Mottaghi et al., 2014)). Since CLIP learns a joint embedding space, text feature similarities directly reflect the model’s ability to distinguish between classes. As illustrated in Figure 2 (top-row), we consistently observe high similarity values between different classes. Specifically, the similarity reaches 0.84 for person-dog pairs and 0.80 for person-horse pairs, far exceeding what one would expect from their semantic relationships. Extending this analysis across various datasets (Table 4), We observe high average feature similarities of 0.77 in VOC, 0.69 in COCO, and 0.75 in Context, indicating that this is a universal limitation of CLIP’s features space. This feature entanglement

fundamentally affects CLIP’s ability to perform fine-grained tasks. When features intended to represent one class encode significant information about other classes, the model struggles to make precise discrimination necessary for tasks like multi-label recognition and semantic segmentation.

To address this limitation, we propose a framework that reduces mutual information between class features while preserving task-essential semantics. Our approach consists of three components: (1) Feature extraction and Projection, where we extract CLIP features and project them into a disentangled space (Section 3.1), (2) Defining novel MFI Loss for disentangling text features (Section 3.2), and (3) Performing MLR to align image features to the disentangled text features (Section 3.3).

3.1. Feature Extraction and Projection

We use CLIP as our feature extractor. Its image encoder ($f_{\theta, \text{img}}$) performs spatial pooling in the final layer, aggregating features from local regions into a d -dimensional vector for the input image x_i . However, this pooling step removes spatial details, making it unsuitable for fine-grained tasks where localization is essential. We remove the final pooling layer to preserve class-specific information across local regions. Then the encoder output for input (\mathbf{x}_i) is $f_{\theta, \text{img}}(\mathbf{x}_i) = \mathbf{z}_i \in \mathbb{R}^{H \times W \times d}$, where H and W are the spatial

dimensions. The text encoder remains unchanged. We use a fixed pair of positive and negative ($\text{txt}_{j,+}$, $\text{txt}_{j,-}$) prompts for each class j as input to the text encoder. The positive prompt indicates the presence of the class in a local region, while the negative prompt indicates its absence. Passing these prompts through the text encoder produces $f_{\theta,\text{text}}(\text{txt}_i) = \mathbf{t}_i \in \mathbb{R}^d$.

The extracted features (image (\mathbf{z}_i), text (\mathbf{t}_i)) lie in CLIP’s original feature space and are not suitable for fine-grained tasks as discussed in Section 1. To address this, we introduce learnable projectors ($h_\phi : h_{\phi,\text{img}}$ and $h_{\phi,\text{text}}$), parameterized by weights ϕ . These projectors map the image (\mathbf{z}_i) and text (\mathbf{t}_i) features from their original space (d -dim) to a new disentangled space (d' -dim), making them suitable for fine-grained tasks. The image projector transforms $\mathbf{z}_i \rightarrow \mathbf{z}'_i$ ($\mathbb{R}^{H \times W \times d} \rightarrow \mathbb{R}^{H \times W \times d'}$) while preserving the spatial dimensions (H, W). The text projector maps $\mathbf{t}_i \rightarrow \mathbf{t}'_i$ ($\mathbb{R}^d \rightarrow \mathbb{R}^{d'}$).

3.2. MFI Loss

We design the projected feature space to reduce mutual feature information (MFI) between class features. Reducing MFI requires obtaining individual class features, as MFI represents the shared information between these individual features. Separating image features into individual class features is non-trivial because multiple classes often co-occur in an image. This leads to mixed features that make class-wise feature isolation difficult. Object segmentation models could assist by extracting features from segmented regions, but these models add significant complexity. In contrast, text class features are inherently independent because they are derived from separate class names or prompts inputted to the text encoder. This independence directly gives us individual text class features. We leverage this property of text features and apply MFI reduction to them.

We propose the MFI reduction loss to minimize the mutual information between class text features. This loss is applied to the projected text features (\mathbf{t}') as follows:

$$\mathcal{L}_{\text{MFI}} = \underbrace{\sum_{i=1} (\mathbf{S}_{ii} - 1)^2}_{\text{Collapse Prevention}} + \lambda \underbrace{\sum_{i=1} \sum_{j=1, j \neq i} \mathbf{S}_{ij}^2}_{\text{MFI Reduction}} \quad (1)$$

where \mathbf{S} is the self-similarity matrix obtained from \mathbf{t}' . Here, \mathbf{S} is defined by

$$\mathbf{S}_{ij} = \frac{\mathbf{t}'_i \mathbf{t}'_j^\top}{\|\mathbf{t}'_i\| \|\mathbf{t}'_j\|}, \quad \forall i, j$$

where $\mathbf{t}'_i, \mathbf{t}'_j$ are the i -th and j -th column vectors of \mathbf{t}' (i.e., $\mathbf{t}'_i, \mathbf{t}'_j \in \mathbb{R}^{d'}$) and $\|\mathbf{t}'_i\|$ is the L_2 -norm of \mathbf{t}'_i . In this formulation, λ is the hyperparameter that addresses the imbalance in the loss arising from the larger number of MFI reduction terms in \mathbf{S} compared to the collapse prevention terms.

The MFI loss minimizes the inter-class similarity \mathbf{S}_{ij} ($i \neq j$) while simultaneously preserving high intra-class \mathbf{S}_{ii} to prevent feature collapse. We provide detailed proof of the loss function’s connection to the Information Bottleneck principle in supplementary material Appendix A.

3.3. Image-Text Alignment with MLR

MLR Formulation. MLR task involves identifying the subset of classes $\mathcal{C}_i \subseteq \{C_1, C_2, \dots, C_N\}$ associated with the image \mathbf{x}_i . The goal is to learn a mapping function $g : \mathbf{x}_i \rightarrow \{-1, 1\}^N$, that maps input images to 1 if the class is present and -1 if the class is absent in the image.

We train our model to recognize multiple objects in images by learning to align projected image features and text features. For each location (h, w) in the projected image features (\mathbf{z}'_i), we detect the presence or absence of a class j , by computing the cosine similarity with positive text features ($\mathbf{t}'_{j,+}$) and negative text features ($\mathbf{t}'_{j,-}$). A higher similarity with the positive text features indicates the presence of the class, while a higher similarity with the negative text features indicates its absence. We aggregate these similarity scores from local regions to produce logits \mathbf{p}_i for the image, following (Sun et al., 2022; Rawlekar et al., 2024b;a). We train the setup with the widely used Asymmetric Loss function (ASL) (Ridnik et al., 2021), which addresses the significant imbalance between negative and positive examples in a multi-label recognition dataset. The ASL loss is given by:

$$\mathcal{L}_{\text{ASL}}(p_i^j) = \begin{cases} \left(1 - p_i^j\right)^{\gamma_+} \log(p_i^j), & \text{if } y_i^j = 1, \\ \left(p_{i,\delta}^j\right)^{\gamma_-} \log(1 - p_{i,\delta}^j), & \text{else} \end{cases} \quad (2)$$

where p_i^j represents the corresponding prediction associated with label y_i^j , i represents the image and j represents the class. $p_{i,\delta}^j = \max(\hat{y} - \delta, 0)$, with δ representing the shifting parameter defined in ASL.

Training. Our training objective is composed of two components: (1) mutual feature information loss that enforces the separation between class text features and (2) Asymmetric loss function (Ridnik et al., 2021), designed for MLR that aligns the image features and text features to obtain predictions for an image.

$$\mathcal{L}_{\text{Unmix-CLIP}} = \mathcal{L}_{\text{ASL}} + \alpha \mathcal{L}_{\text{MFI}} \quad (3)$$

where α controls the relative importance of the two objectives.

4. Experiments

Here we describe the datasets, evaluation metrics, implementation details, and performance analysis for multi-label recognition and zero-shot semantic segmentation.

Table 1: **Comparison on multi-label recognition (MLR).** We compare the performance (mAP) and training efficiency (number of parameters) of our approach with SOTA VLM-based MLR methods on VOC2007 and COCO-14 datasets. Our approach is competitive with SOTA on VOC2007, and on the challenging COCO dataset, it outperforms SOTA while requiring only one-third of the parameters. **red** and **blue** indicate the best and the second best performance.

Methods	VOC2007		COCO-14	
	# Params(\downarrow)	mAP(\uparrow)	# Params (\downarrow)	mAP(\uparrow)
DualCoOp (Sun et al., 2022)	0.3M	94.2	1.3M	83.6
SCPNet (Ding et al., 2023)	-	94.3	3.4M	84.4
TAI-DPT (Guo et al., 2023)	> 0.3M	-	>1.3M	84.5
DualCoOp++ (Hu et al., 2023)	0.4M	94.9	1.5M	85.1
MLR-GCN (Rawlekar et al., 2024b)	0.3M	94.4	1.3M	-
PositiveCoOp (Rawlekar et al., 2024a)	0.2M	94.4	0.8M	84.7
Ours	0.4M	94.8	0.4M	85.3

4.1. Datasets and Metrics

1) Pre-training with MLR: We evaluate the MLR performance using mean-Average Precision (mAP) on the following datasets:

COCO-14 (Lin et al., 2014) contains 80 classes across diverse categories with 82,081 training and 40,504 validation images. Following recent works (Sun et al., 2022; Rawlekar et al., 2024b;a), we train on the training set and evaluate on the validation set.

VOC2007 (Everingham et al., 2010) is another widely used MLR dataset containing 20 classes with 9,963 images. Following (Sun et al., 2022; Rawlekar et al., 2024b;a), we use the train-val set for training and the test set for evaluation.

2) Zero-Shot Semantic Segmentation (ZS3): We use image and text projectors trained on the COCO-14 dataset and evaluate ZS3 using the mIoU metric on the following datasets:

PASCAL VOC 2012 (Everingham et al., 2010) includes segmentation masks for the 20 classes in VOC2007. Following works (Li et al., 2023; Bousselham et al., 2024), we evaluate on the validation set.

PASCAL Context (Mottaghi et al., 2014) extends PascalVOC to 59 classes, 30 of which were unseen during our pre-training. These additional classes provide dense annotations for the whole scene. We evaluate the test set, comprising 5,104 images.

COCO-2017 (Lin et al., 2014) includes segmentation masks for the 80 classes in COCO-14. Following (Li et al., 2023; Bousselham et al., 2024), we evaluate the validation set.

4.2. Implementation Details

We use CLIP’s (Radford et al., 2021) original pre-trained encoder weights for all our experiments and keep them frozen. Consistent with popular MLR and ZS3 literature, we use a ResNets-based visual encoder (RN-101) and the standard transformer for text encoding (Sun et al., 2022; Ding et al., 2023; Hu et al., 2023; Rawlekar et al., 2024a;b;

Guo et al., 2023; Li et al., 2023; Lin et al., 2023). We conduct all experiments on a single RTX A4000 GPU.

During the pre-training stage with the MLR setup (Section 3.3), we follow the settings and hyperparameters from recent works (Sun et al., 2022; Rawlekar et al., 2024b;a). This includes resizing images to 448, applying Cutout (DeVries & Taylor, 2017) and RandAugment (Cubuk et al., 2020) for augmentation. Our projectors (h_ϕ) are implemented as multi-layer perceptrons (MLPs). Specifically, the image projector follows a $[512 \rightarrow 256]$ architecture, while the text projector is designed as $[512 \rightarrow 384 \rightarrow 256]$ with batch normalization and ReLU. We train both projectors with stochastic gradient descent (SGD) using an initial learning rate of 0.002, which is reduced by cosine annealing. We train the Unmix-CLIP setup (ASL + MFI loss) for 50 epochs with a batch size of 32. We follow (Sun et al., 2022; Rawlekar et al., 2024b;a), and use ASL hyperparameters in Equation (2) as $\gamma_- = 2$, $\gamma_+ = 1$ and $\delta = 0.05$. We set $\lambda = 0.2$ and $\alpha = 7e-5$ when pre-trained with COCO-14 in Equation (1).

For Zero-Shot Semantic Segmentation, we adopt the v-v attention described in (Li et al., 2023) that prevents inversion of activation commonly observed in CLIP. We then add our pre-trained projectors to CLIP. To obtain the segmentation mask, we compute the cosine similarity between locally projected image features (\mathbf{z}') and projected text features for all classes in the dataset. We use the text template “A photo of a {classname}.” Lastly, we use bilinear interpolation to upsample the segmentation mask to the input image size.

4.3. Results

Multi-Label Recognition. We primarily compare Unmix-CLIP with other SOTA VLM-based MLR approaches. In Table 1, we present a detailed comparison of the performance (mAP) and the number of training parameters required by each method on the VOC2007 (Everingham et al., 2010)

Table 2: **Comparison on zero-shot semantic segmentation (ZS3).** We compare Unmix-CLIP with other SOTA baselines across three semantic segmentation datasets using the mIoU metric. The "Dataset" column details the pre-training dataset and the type of annotations used. The abbreviations are as follows: Loc Ann. + FT: local annotations and fine-tuning, SM: segmentation mask, IT: image-text, IC: image classes, Bkgd: include background class, No Bkgd: ignore background class, MR: MFI Reduction(%), red and blue indicate the best and the second best performance

Method	Loc Ann.	Dataset		VOC12		COCO-17		Context
		Pre-training	Ann	Bkgd	Bkgd	No Bkgd		
SPNet(Xian et al., 2019)	✓	COCO, VOC, Context	SM	15.6	-	-	4	
ZS3Net(Bucher et al., 2019)	✓	VOC, Context	SM	17.7	-	-	7	
CLIP-ES(Lin et al., 2023)	✓	WIT, COCO-Stuff	IT,IC	75	-	-	-	
CLIP(Radford et al., 2021)	✗	WIT-400M	IT	14.1	3.9	5.6	4.1	
CLIPSurgery(Li et al., 2023)	✗	WIT-400M	IT	17.5	13.0	22.9	11	
CLIP-VV(Li et al., 2023)	✗	WIT-400M	IT	<u>32.6</u>	<u>19.9</u>	<u>35.5</u>	<u>15.5</u>	
Ours (MR = 24.9)	✗	WIT-400M, COCO	IT,IC	36	22.7	37.8	12.9	

and COCO-14 (Lin et al., 2014) datasets. For VOC 2007, we observe that our performance is competitive with DualCoOp++(Hu et al., 2023) and requires the same number of parameters. However, on the more challenging COCO-14 dataset, Unmix-CLIP outperforms DualCoOp++ while requiring only one-third of the training parameters.

Zero-Shot Semantic Segmentation. We categorize our comparisons into two main groups. The first group includes approaches that use local annotations (segmentation masks, etc.) to fine-tune the network (Xian et al., 2019; Bucher et al., 2019; Lin et al., 2023). The second comparison is with training-free approaches (Radford et al., 2021; Li et al., 2023). Our approach is closer to the training-free methods, as it does not use any form of local annotations.

Our results are summarized in Table 2. Following prior works (Xian et al., 2019; Bucher et al., 2019; Lin et al., 2023; Li et al., 2023), we report mIoU values for VOC 2012 by including the background as a class. We use a threshold of 0.85 to identify the background, as suggested in (Bousselham et al., 2024). Our approach outperforms CLIP Surgery by 18.5 mIoU and CLIP-VV by 3.4 mIoU on VOC 2012. For COCO-14, we report results both with and without the background class. When including the background, our method surpasses CLIP Surgery and CLIP-VV by 9.7 mIoU and 2.8 mIoU, respectively. Without the background, we achieve gains of 14.9 mIoU and 2.3 mIoU. Additionally, we evaluate the VOC Context dataset, which contains 30 unseen classes not used during our pre-training. Although our model is not explicitly trained to reduce MFI between these classes (it is designed to minimize MFI among COCO’s 80 classes), our approach still outperforms CLIP Surgery. These results demonstrate that our projectors preserve some of the open-vocabulary capabilities of CLIP. We show qualitative results for open-vocabulary tasks in Supplementary Figure 7.

Table 3: **Comparison on Zero-shot Multi-Label Recognition (ZS-MLR).** We segment objects from images using Unmix-CLIP and improve CLIP’s zero-shot multi-label recognition capabilities by integrating predictions from segmented objects and the entire image.

Dataset	Backbone	CLIP (mAP)	Ours (mAP)
VOC2007	RN 101	78.73	80.71
	RN 50	76.20	79.87
COCO-14	RN 101	50.10	52.00
	RN 50	47.30	50.15

Segmentation-driven Zero-Shot Multi-Label Recognition (ZS-MLR). We leverage the segmentation capabilities of Unmix-CLIP to reformulate the multi-label recognition problem into a single-label recognition problem, a domain more suitable for CLIP. Specifically, we use two predictions: global and local. We pass the input image directly through CLIP to obtain its global predictions. However, as discussed in Section 1, these predictions are often dominated by more prominent objects in the image, ignoring smaller objects, which leads to poor zero-shot MLR performance. To address this limitation, we introduce local predictions. We segment the image into multiple regions (ideally corresponding to individual objects) using Unmix-CLIP. Each segment is then processed independently through CLIP, and the predictions from all segments are aggregated. Finally, we combine the global and local predictions to obtain the zero-shot scores for the image. We evaluate the ZS-MLR performance on the VOC2007 and COCO-14 datasets, with results presented in Table 3. The results demonstrate that our method provides meaningful information (segments) to improve CLIP zero-shot capabilities.

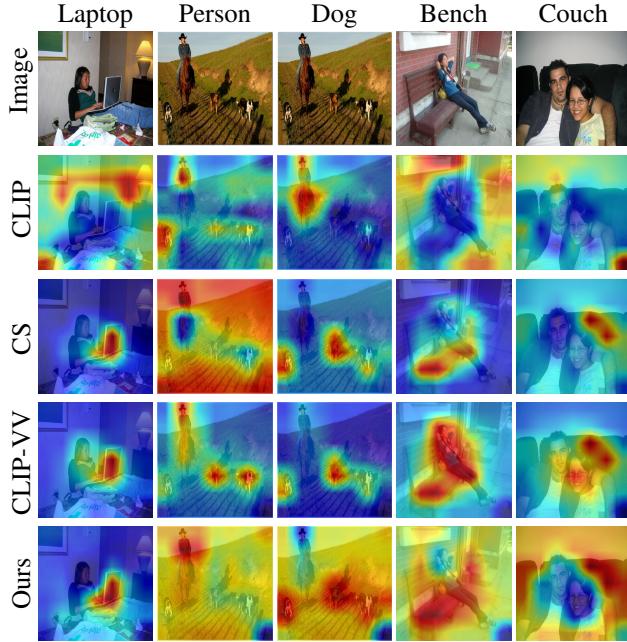


Figure 4: **Qualitative Comparison on ZS3.** Visualization of zero-shot semantic segmentation (ZS3) results for CLIP (Radford et al., 2021), CLIP Surgery (CS) (Li et al., 2023), CLIP-VV (Li et al., 2023), and our approach across multiple categories. The heatmaps show activation regions for each queried class, where darker red indicates strongly activated regions. Our method produces more separated activations, demonstrating improved class localization.

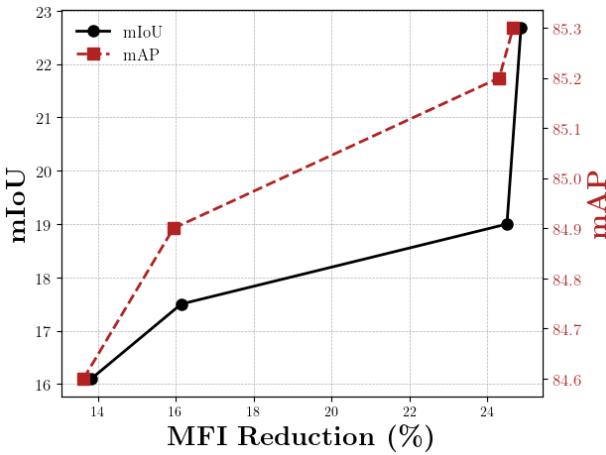


Figure 5: **Performance vs. MFI Reduction.** Performance of Multi-Label Recognition (MLR, measured by mAP) and Zero-Shot Semantic Segmentation (ZS3, measured by mIoU) on COCO as a function of MFI reduction. As class feature separation increases (i.e., MFI decreases), the model performs better on both tasks.

Table 4: **Quantitative MFI Reduction.** MFI values are reported across different dataset datasets with seen (VOC, COCO), partial (Context), and unseen (Context) classes. Our method significantly reduces MFI for all datasets.

Method	VOC	COCO	Context	
	Seen	Seen	Partial	Unseen
CLIP	0.77	0.69	0.75	0.75
Ours	0.50	0.52	0.53	0.52
Δ (%)	34.8	24.9	29.8	30.4

Table 5: **MFI Loss Ablation Study.** Adding MFI Loss to our method improves multi-label recognition (MLR) performance by 0.5 mAP on the COCO-14 dataset, demonstrating its effectiveness for fine-grained tasks.

Method	ASL Loss	MFI Loss	mAP
Ours	✓	✗	84.8
Ours	✓	✓	85.3

5. Analysis

Feature Disentanglement. We pre-train Unmix-CLIP on COCO-14, which contains 80 classes. As shown in Section 4.3, our approach improves performance even on datasets with previously unseen classes, such as VOC Context. We analyze this improvement by comparing MFI reduction across four datasets: VOC2012 (20 seen classes), COCO-2017 (80 seen classes), Context (59 partially seen classes), and a Context subset (30 unseen classes from COCO-2017). Figure 2 shows the self-similarity matrices of class text features from CLIP and Unmix-CLIP, demonstrating the class feature disentanglement. Table 4 quantifies the MFI reduction through the difference in average inter-class similarity between CLIP and Unmix-CLIP. Our framework effectively disentangles representations for both seen and unseen classes, leading to performance gains.

Feature Disentanglement Impact. Figure 5 shows how MFI reduction improves performance in multi-label recognition on COCO-14 dataset and zero-shot semantic segmentation on the COCO-2017 dataset. We observe that as MFI decreases, the performance of both tasks improves.

6. Conclusions

In conclusion, this work advances our understanding of CLIP features by identifying and addressing a fundamental challenge in their localized understanding. We first show that reducing mutual information is critical for fine-grained recognition tasks. Motivated by this, we introduce Unmix-CLIP, a novel approach to project CLIP features into a disentangled space by combining our proposed MFI loss and

the asymmetric loss for MLR. Our experimental results demonstrate that reducing feature entanglement through Unmix-CLIP significantly enhances the model’s ability to perform fine-grained tasks. This improvement is particularly evident in two challenging tasks: multi-label recognition (MLR) and zero-shot semantic segmentation (ZS3). These findings highlight the importance of feature disentanglement in vision-language models and provide a promising direction for future research in improving the localized understanding capabilities of CLIP-based architectures. A limitation of our approach is its reduced capability in zero-shot open-vocabulary segmentation. We constrain some of CLIP’s broader semantic capabilities by optimizing feature disentanglement for COCO dataset classes. Training on substantially larger datasets could help mitigate this limitation while preserving the benefits of our feature disentanglement approach.

References

- Baldrati, A., Bertini, M., Uricchio, T., and Del Bimbo, A. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21466–21474, 2022.
- Ballé, J., Laparra, V., and Simoncelli, E. P. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016.
- Barlow, H. B. et al. Possible principles underlying the transformation of sensory messages. *Sensory Communication*, 1(01):217–233, 1961.
- Bousselham, W., Petersen, F., Ferrari, V., and Kuehne, H. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3828–3837, 2024.
- Bucher, M., Vu, T.-H., Cord, M., and Pérez, P. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019.
- Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020.
- Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.
- Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 702–703, 2020.
- Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.
- DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.
- Ding, Z., Wang, A., Chen, H., Zhang, Q., Liu, P., Bao, Y., Yan, W., and Han, J. Exploring structured semantic prior for multi label recognition with incomplete labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3398–3407, 2023.
- Dong, X., Bao, J., Zheng, Y., Zhang, T., Chen, D., Yang, H., Zeng, M., Zhang, W., Yuan, L., Chen, D., et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10995–11005, 2023.
- Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88: 303–338, 2010.
- Guo, Z., Dong, B., Ji, Z., Bai, J., Guo, Y., and Zuo, W. Texts as images in prompt tuning for multi-label image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2808–2817, 2023.
- He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.
- Henaff, O. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pp. 4182–4192. PMLR, 2020.
- Hu, P., Sun, X., Sclaroff, S., and Saenko, K. Dualcoop++: Fast and effective adaptation to multi-label recognition with limited annotations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- Huang, H., Rawlekar, S., Chopra, S., and Deniz, C. M. Radiology reports improve visual representations learned from radiographs. In *Medical Imaging with Deep Learning*, pp. 1385–1405. PMLR, 2024.
- Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with

- noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.
- Li, Y., Wang, H., Duan, Y., and Li, X. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.
- Lin, Y., Chen, M., Wang, W., Wu, B., Li, K., Lin, B., Liu, H., and He, X. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15305–15314, 2023.
- Lin, Y., Chen, M., Zhang, K., Li, H., Li, M., Yang, Z., Lv, D., Lin, B., Liu, H., and Cai, D. Tagclip: A local-to-global framework to enhance open-vocabulary multi-label classification of clip without training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3513–3521, 2024.
- Lüddecke, T. and Ecker, A. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7086–7096, 2022.
- Mottaghi, R., Chen, X., Liu, X., Cho, N.-G., Lee, S.-W., Fidler, S., Urtasun, R., and Yuille, A. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 891–898, 2014.
- Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.
- Rawlekar, S., Bhatnagar, S., and Ahuja, N. Rethinking prompting strategies for multi-label recognition with partial annotations. *arXiv preprint arXiv:2409.08381*, 2024a.
- Rawlekar, S., Bhatnagar, S., Srinivasulu, V. P., and Ahuja, N. Improving multi-label recognition using class co-occurrence probabilities. *International Conference on Pattern Recognition*, 2024b.
- Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Proter, M., and Zelnik-Manor, L. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 82–91, 2021.
- Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.
- Sun, X., Hu, P., and Saenko, K. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *Advances in Neural Information Processing Systems*, 35: 30569–30582, 2022.
- Tishby, N. and Zaslavsky, N. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pp. 1–5. IEEE, 2015.
- Xian, Y., Choudhury, S., He, Y., Schiele, B., and Akata, Z. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8256–8265, 2019.
- Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.
- Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.
- Zhou, C., Loy, C. C., and Dai, B. Extract free dense labels from clip. In *European Conference on Computer Vision*, pp. 696–712. Springer, 2022a.
- Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.
- Zhu, K. and Wu, J. Residual attention: A simple but effective method for multi-label recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 184–193, 2021.

A. Objective Function: MFI Loss

This section establishes a connection between MFI loss and the Information Bottleneck (IB) principle (Tishby & Zaslavsky, 2015). As described in Section 3.2, MFI loss explicitly reduces the mutual information between text features to obtain disentangled features.

A.1. Information Bottleneck (IB) Objective

Formulation. Let, T_i represent the input text (i.e., the prompt with the class label), and let Z_i be the extracted features from the CLIP text encoder (Radford et al., 2021). The output is represented by Y_i , indicating the class associated with Z_i .

As we show in Section 3, mutual information exists between text (class) features Z_i , i.e., each class feature contains information about multiple classes rather than only its corresponding class Y_i . Our goal is to enforce a one-to-one mapping where Z_i retains information only about Y_i while discarding information about all other classes Y_j ($j \neq i$).

This aligns naturally with the IB principle, which formulates an optimal trade-off between minimizing the information Z_i retains from T_i and maximizing the information it preserves for the target class Y_i . We extend the IB principle to reduce explicit information about all other classes. We express this formally as:

$$\text{IB} = \mathbf{I}(Z_i, T_i) + \beta \left[\mathbf{I}(Z_i, Y_i) - \sum_{j \neq i} \mathbf{I}(Z_i, Y_j) \right] \quad (4)$$

where \mathbf{I} represents mutual information. Here,

1. $I(Z_i; T_i)$ ensures that Z_i take only the information from T_i that is needed to map to Y_i .
2. $I(Z_i; Y_i)$ preserves discriminative class information.
3. $\sum_{j \neq i} I(Z_i; Y_j)$ reduces information in Z_i that map to Y_j where $j \neq i$

A.2. Connection to MFI Loss

To minimize IB, we first express mutual information in terms of entropy:

$$\mathbf{I}(A; B) = \mathbf{H}(A) - \mathbf{H}(A|B), \quad (5)$$

where $\mathbf{H}(A)$ is the marginal entropy of A , and $\mathbf{H}(A|B)$ is the conditional entropy of A given B .

Substituting this into the IB objective Equation (4):

$$\begin{aligned} \text{IB} = & (1 + \beta - \sum_{j \neq i} \beta) \mathbf{H}(Z_i) - \beta \left[\mathbf{H}(Z_i|Y_i) - \sum_{j \neq i} \mathbf{H}(Z_i|Y_j) \right] \\ & - \mathbf{H}(Z_i|T_i) \end{aligned} \quad (6)$$

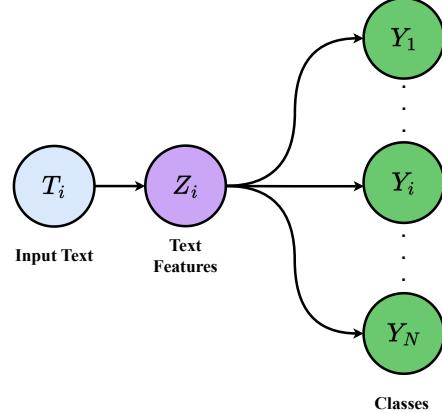


Figure 6: The Information Bottleneck principle is applied for feature disentanglement. Given an input text T_i , the text encoder of CLIP (Radford et al., 2021) generates features Z_i , which encode information about the output classes Y_i . Our objective is to ensure that Z_i retains only the information necessary to map to its corresponding class Y_i while minimizing its information about other classes Y_j ($j \neq i$)

Since the CLIP text encoder is deterministic, the entropy term $\mathbf{H}(Z_i|T_i) = 0$. Also, given that text inputs are predefined (i.e., class names in the dataset), Z_i is deterministic, implying $\mathbf{H}(Z_i) = 0$. This simplifies the IB objective to:

$$\text{IB} \propto -\mathbf{H}(Z_i|Y_i) + \sum_{j \neq i} \mathbf{H}(Z_i|Y_j). \quad (7)$$

Assuming Z follows a Gaussian distribution, its entropy is given by:

$$\mathbf{H}(Z) = \frac{1}{2} \log |\mathbf{C}| + \text{const}, \quad (8)$$

where \mathbf{C} is the covariance matrix of Z_i . Since the constant term does not affect the optimization, we optimize the determinant of the covariance matrix \mathbf{C} . In practice, we optimize the covariance matrix. Thus, minimizing IB reduces the covariance between class features, ensuring they are independent.

The IB objective in Equation (7) becomes:

$$\text{IB} \propto -\mathbf{C}_{Z_i|Y_i} + \sum_{j \neq i} \mathbf{C}_{Z_i|Y_j}, \quad (9)$$

Minimizing the IB objective is equivalent to minimizing the MFI Loss. Specifically, maximizing $\mathbf{C}_{Z_i|Y_i}$ is equivalent to collapse prevention term and minimizing $\sum_{j \neq i} \mathbf{C}_{Z_i|Y_j}$ is our MFI reduction term in the following equation:

$$\mathcal{L}_{\text{MFI}} = \underbrace{\sum_{i=1} (\mathbf{S}_{ii} - 1)^2}_{\text{Collapse Prevention}} + \lambda \underbrace{\sum_{i=1} \sum_{j=1, j \neq i} \mathbf{S}_{ij}^2}_{\text{MFI Reduction}} \quad (10)$$



Figure 7: **Qualitative comparison - Open Vocabulary.** Comparison of segmentation results of Unmix-CLIP on various unseen classes, including fine-grained categories such as celebrities and animated characters.

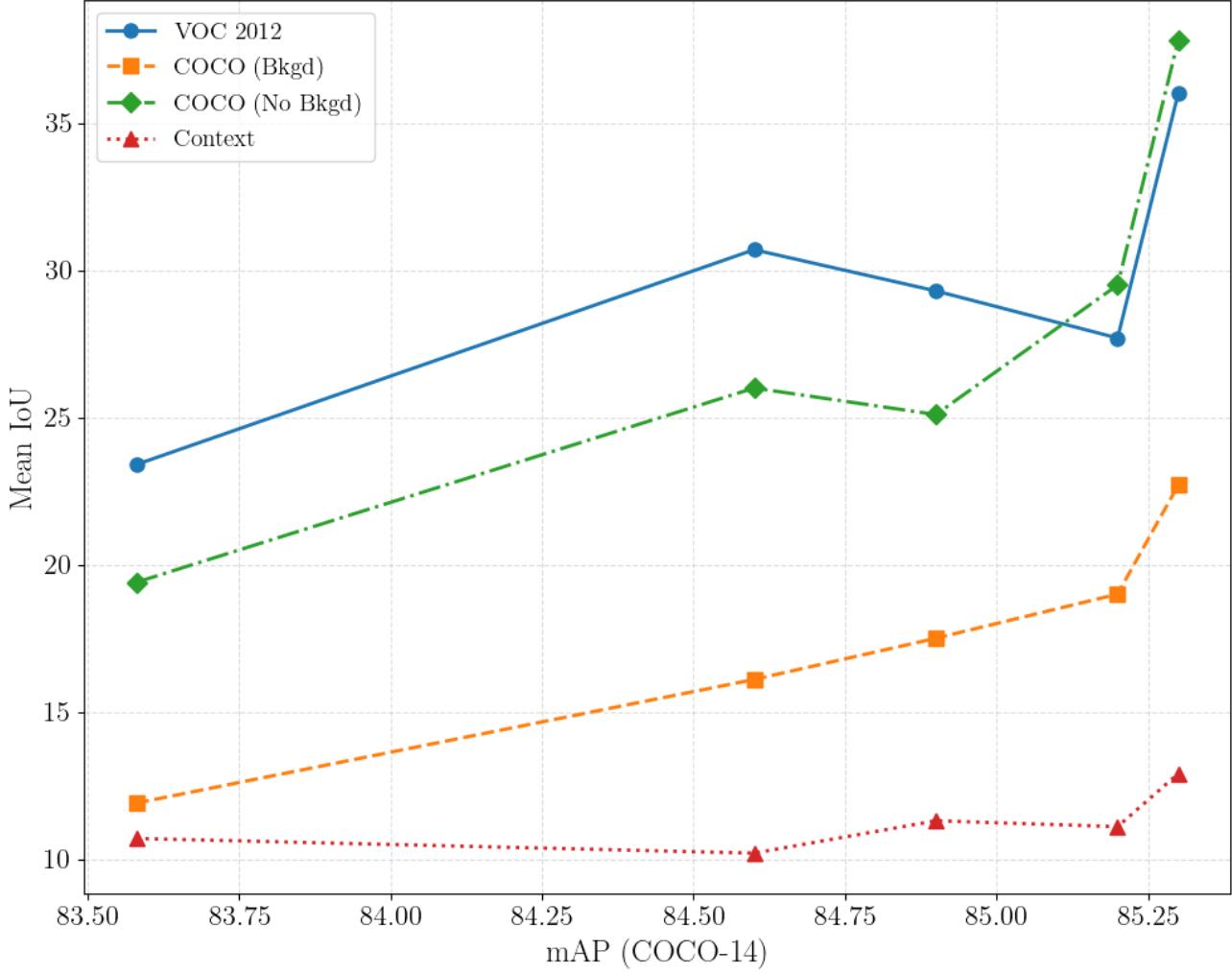


Figure 8: **mAP vs mIoU.** Performance comparison of zero-shot semantic segmentation (mIoU) for VOC2012, COCO 2017 with and without the background, and VOC Context as a function of multi-label recognition (mAP) performance on the COCO-14 dataset. A general trend: higher MLR performance positively correlates with segmentation results.