DETAILCLIP:
DETAIL-ORIENTED CLIP FOR FINE-GRAINED TASKS

Amin Karimi Monsefi[§*]   **Kishore Prakash Sailaja**[§*]   **Ali Alilooee**[§]
Ser-Nam Lim[‡]   **Rajiv Ramnath**[§]
{karimimonsefi.1, prakashsailaja.1, alilooeedolatabad.1,
ramnath.6}@osu.edu, sernam@ucf.edu
[§]The Ohio State University,   [‡]University of Central Florida

ABSTRACT

In this paper, we introduce DetailCLIP, a self-improving vision-language foundation model designed to enhance fine-grained feature understanding through self-supervised learning. Foundation models like CLIP have demonstrated strong performance in global image-text alignment but often fail to capture detail-oriented features necessary for tasks such as segmentation. To address this, DetailCLIP integrates self-curated learning objectives that iteratively improve both high-level semantics and detailed visual representations. Specifically, our method employs patch-level self-distillation and pixel-level reconstruction losses to generate refined internal representations, while an attention-based token filtering mechanism curates semantically relevant information during training. By generating and refining self-curated learning signals, DetailCLIP improves segmentation performance and demonstrates superior generalization across diverse tasks. These task-agnostic objectives position DetailCLIP as a self-improving foundation model, enhancing multi-modal systems like CLIP with fine-grained feature understanding. https://github.com/KishoreP1/DetailCLIP

1   INTRODUCTION

The rapid advancements in computer vision have led to the development of foundation models that can understand visual data with impressive accuracy. Among these, CLIP (Contrastive Language-Image Pretraining) (Radford et al., 2021) stands out as a pioneering approach that leverages large-scale contrastive learning between images and text to create a shared embedding space. This has proven immensely successful in tasks like classification, where the model can understand and relate visual content to textual descriptions without direct supervision (Yu et al., 2024; Wei et al., 2023; Lan et al., 2024).

However, while CLIP excels in generalization and high-level semantic understanding, its reliance on contrastive losses presents significant challenges when adapting the model for more granular tasks like image segmentation (Wei et al., 2023; Zhou et al., 2023b; Lin et al., 2023; Yang et al., 2023). Fine-grained tasks like segmentation requires a comprehensive understanding of the entire scene and the precise delineation of object boundaries at the pixel level. However, the global representations learned through contrastive learning often fail to capture this level of detail, leading to suboptimal performance when using CLIP for granular tasks (Jiao et al., 2023; Cha et al., 2023; Zhou et al., 2023a; Karimi Monsefi et al., 2023; Perera et al., 2024; Wang et al., 2022).

Furthermore, traditional self-supervised learning (SSL) approaches (Zhou et al., 2021; Caron et al., 2021; Kakogeorgiou et al., 2022; Zhou et al., 2024; Navard & Yilmaz, 2024; Farahani & Monsefi, 2023), which have gained popularity for their ability to learn from unlabeled data, often prove inadequate when applied to fine-grained tasks. These methods, including contrastive learning (Oord et al., 2018; Chen et al., 2020a; Tian et al., 2020; He et al., 2020) or clustering-based techniques (Monsefi et al., 2024), primarily focus on learning representations that are useful for high-level tasks such as classification or detection. However, they often fail to capture the intricate details and spatial

---

[*]Equal contribution.

relationships necessary for precise boundary delineation and detailed feature extraction. This limitation is particularly evident in masked image modeling (MIM) approaches such as MAE (He et al., 2022), SimMIM (Xie et al., 2022), and iBOT (Zhou et al., 2021), which, while effective at learning global features, struggle to preserve fine-grained details. Consequently, these models often require extensive training epochs to achieve moderate performance improvements, yet still fall short in tasks that demand a high degree of granularity. The lack of textual context in these SSL models further exacerbates their limitations, as they miss the additional semantic guidance that text provides, which is crucial for enhancing the model's ability to focus on detail-oriented tasks. Incorporating text in the pre-training process can significantly improve the model's performance by aligning visual features with rich semantic information, leading to better generalization and more accurate granular tasks (Bao et al., 2022; Yi et al., 2023; Zakeri et al., 2019).

This underscores the need for self-curated learning mechanisms that provide stronger, task-agnostic signals, enabling models to refine both high-level and fine-grained representations. To address this, we introduce DetailCLIP, a self-improving framework that overcomes the limitations of traditional CLIP-based models and self-supervised approaches for detail-oriented tasks. By enhancing both semantic understanding and fine-grained feature extraction, DetailCLIP achieves superior performance in tasks requiring meticulous attention to detail.

DetailCLIP introduces an attention-based token removal mechanism that selectively retains the most relevant image regions, guided by task-specific and textual information. As shown in Figure 3, this mechanism focuses the model on critical areas, enhancing its ability to capture and represent subtle details accurately in the output.

DetailCLIP improves pixel-level accuracy through a pixel-level reconstruction feature that preserves detailed image features, enabling it to handle fine-grained tasks requiring precision. Unlike models like MAE (He et al., 2022) and SimMIM (Xie et al., 2022), which use random token removal, DetailCLIP employs an attention-based token removal mechanism to retain only the most relevant features. This targeted approach enhances performance on tasks demanding high pixel-level accuracy, setting DetailCLIP apart from previous methods.

DetailCLIP incorporates a self-distillation strategy (Caron et al., 2021), where the student model reconstructs masked images under the guidance of a teacher model using unmasked images. This iterative process refines intricate features, enhancing both segmentation accuracy and performance on classification tasks by preserving semantic integrity and detailed feature recognition. Unlike previous models such as SLIP (Mu et al., 2022), MaskCLIP (Dong et al., 2023), and A-CLIP (Yang et al., 2023), which focus on coarse-grained comparisons, DetailCLIP combines both coarse-grained and fine-grained self-distillation, enabling superior performance in tasks requiring high precision.

Finally, similar to CLIP-based models (Radford et al., 2021; Luo et al., 2023), DetailCLIP emphasizes the crucial connection between vision and language, which is essential for tasks that require a deep understanding of visual content and its associated textual concepts.

To summarize, our contributions are fourfold:

- A **self-improving token filtering mechanism** driven by attention, which selectively retains tokens with strong semantic relevance to both textual descriptions and fine-grained tasks.
- A **patch-level self-distillation strategy**, where the model learns from masked images by comparing internal representations.
- A **pixel-level reconstruction** of masked images to provide a stronger self-supervised learning signal, enabling accurate recovery of intricate details for precision tasks.
- Integration of these objectives makes DetailCLIP an effective **self-improving vision-language model**, excelling in zero-shot and fine-tuning for detail-oriented tasks.

2 RELATED WORK

2.1 SELF SUPERVISED LEARNING

Self-Supervised Learning (SSL) has emerged as a dominant force in visual pre-training, with various strategies driving its success across different pretext tasks (Ben-Shaul et al., 2024; Su et al., 2024;

Almalki & Latecki, 2024; Oord et al., 2018; Chen et al., 2020a; Tian et al., 2020; He et al., 2020; Bao et al., 2022; He et al., 2022; Xie et al., 2022; Monsefi et al., 2024). These approaches enable SSL to extract meaningful visual features from large-scale unlabeled image datasets (Oquab et al., 2024; Xie et al., 2023; Liu et al., 2023; Zheng et al., 2024; Kakogeorgiou et al., 2022; Zhou et al., 2021; Chen et al., 2020b; Chen & He, 2021; Caron et al., 2021). Among the key methodologies are contrastive learning, Masked Image Modeling (MIM), Masked Frequency Modeling (MFM), and self-supervised Knowledge Distillation (KD), each designed to optimize representation learning in different ways.

Contrastive learning, a popular SSL technique, aligns different views of the same image within a shared embedding space while distinguishing them from views of other images (Oord et al., 2018; Chen et al., 2020a). Although this approach is effective for learning global features, it often faces challenges in handling tasks that require detailed precision, such as image segmentation. The reason is that contrastive learning primarily focuses on aligning overall image representations, which can lead to the neglect of fine-grained, pixel-level details.

On the other hand, Masked Image Modeling (MIM) has shown greater promise for tasks requiring detailed feature extraction. MIM involves masking portions of an image and training models to reconstruct or predict the missing parts, encouraging the model to understand the intricate relationships between visible and hidden regions (Bao et al., 2022; Xie et al., 2022; He et al., 2022; Yi et al., 2023).

2.2 CLIP-BASED APPROACHES

The integration of self-supervised learning with vision-language pre-training has driven the development of several CLIP-based models, each striving to bridge the gap between visual and textual data (Cha et al., 2023; Zhou et al., 2023a; Li et al., 2023; Jia et al., 2021). These models build on the original CLIP framework, introducing innovations to enhance the quality of representations and improve alignment between images and text.

For instance, SLIP (Mu et al., 2022) enhances CLIP by incorporating self-supervised learning with image-to-image contrastive learning, resulting in richer and more robust visual representations. MaskCLIP (Dong et al., 2023) advances this concept by integrating masked image modeling, which refines visual features by focusing on specific image regions, ensuring they are more closely aligned with the accompanying text. A-CLIP (Yang et al., 2023) further refines this approach by implementing an attentive token removal strategy, selectively preserving tokens that are semantically relevant to the text, thereby enhancing the precision of the visual-text alignment.

While these methods enhance the original CLIP model by improving representation learning and efficiency, they primarily focus on global alignment between images and text. This emphasis on global features makes them less effective for detail-oriented tasks, where capturing and preserving fine-grained visual details is crucial (Jiao et al., 2023; Wei et al., 2023; He et al., 2023; Zhou et al., 2023b). In many of these models, token removal is often random or solely based on textual information (Dong et al., 2023; Yang et al., 2023). In contrast, DetailCLIP addresses these limitations by employing an attention-based mechanism that incorporates both textual information and fine-grained tasks, ensuring better performance in detail-sensitive applications.

3 METHOD

3.1 PRELIMINARY AND BACKGROUND

CLIP, developed by Radford et al. (2021), is a vision-language model that aligns visual and textual representations through large-scale contrastive learning. It trains an image and text encoder jointly to maximize similarity between matching image-text pairs and minimize it for mismatched ones. The training objective uses a symmetric cross-entropy loss on the cosine similarity between image and text embeddings, encouraging correct pairs to have higher similarity than incorrect pairs. Let I_i and T_j represent the embeddings of the i-th image and j-th text in a batch. The cosine similarity between them is defined as:

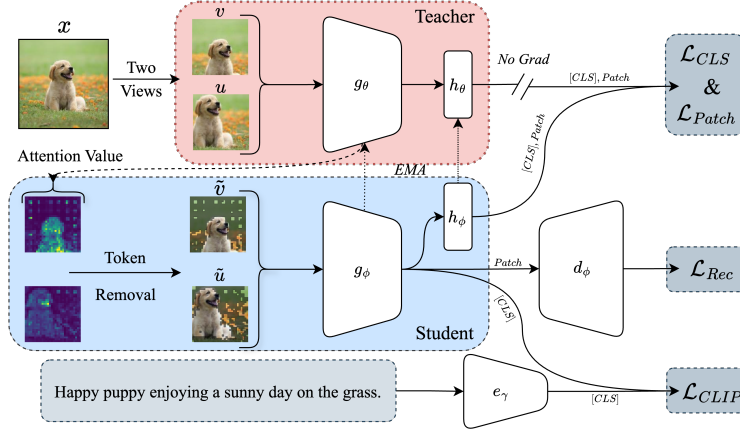$$s_{i,j} = \frac{I_i \cdot T_j}{\|I_i\|\|T_j\|}, \tag{1}$$

Figure 1: Using a teacher-student architecture, the teacher processes two views of an input image to generate attention values, guiding token removal in the student model. The student reconstructs the image with a vision decoder while optimizing classification loss ($\mathcal{L}_{CLS}$), patch contrastive loss ($\mathcal{L}_{Patch}$), and reconstruction loss ($\mathcal{L}_{Rec}$). The CLIP loss ($\mathcal{L}_{CLIP}$) ensures alignment between vision and text encoders, enhancing segmentation. In the figure, g is the vision encoder, h the projection head, d the decoder, and e the text encoder.

The logits are scaled by a learnable parameter τ:

$$\text{logits}_{i,j} = \frac{s_{i,j}}{\tau}, \tag{2}$$

Let N be the number of pairs of image-text in a batch. The loss function for the image-to-text matching is defined as:

$$\mathcal{L}_{I2T} = -\frac{1}{N} \sum_{i=1}^{N} \log \left(\frac{\exp(\text{logits}_{i,i})}{\sum_{j=1}^{N} \exp(\text{logits}_{i,j})} \right), \tag{3}$$

Similarly, the loss function for the text-to-image matching is:

$$\mathcal{L}_{T2I} = -\frac{1}{N} \sum_{i=1}^{N} \log \left(\frac{\exp(\text{logits}_{i,i})}{\sum_{j=1}^{N} \exp(\text{logits}_{j,i})} \right), \tag{4}$$

Finally, the overall loss for CLIP is the average of the two losses:

$$\mathcal{L}_{CLIP} = \frac{1}{2}(\mathcal{L}_{I2T} + \mathcal{L}_{T2I}), \tag{5}$$

3.2 DETAILCLIP FRAMEWORK

As shown in Figure 1, our framework can be outlined through the following steps, which will be discussed in greater detail. Additionally, Section A.3 of the appendix provides a pytorch-like pseudocode of the forward pass.

3.2.1 PATCH-LEVEL COMPARISON

We employ a teacher-student framework in which the student model is trained to predict both fine-grained and coarse-grained features of images generated by the teacher model. The teacher model provides "target" features for the student to learn. For each input, we consider two views: the

original views are fed into the teacher model, while their masked versions are used as input for the student model. For further details on the masking process, refer to Section 3.2.3. The student model is tasked with predicting both the masked tokens and the global features generated by the teacher.

Similar to methods used in previous works like DINO (Caron et al., 2021), the teacher model's parameters are updated using an exponential moving average (EMA) of the student model's parameters and its own. Specifically, if θ represents a parameter of the teacher model and ϕ represents a parameter of the student model, the update rule for θ is $\theta \leftarrow (1 - \lambda)\phi + \lambda\theta$, that λ starting at 0.996 and gradually increasing to 1.0, following a cosine scheduling approach as outlined in previous works (Chen et al., 2021; Dong et al., 2023; Yang et al., 2023). This strategy ensures that the EMA model remains stable and effective in capturing the most relevant features during training (Grill et al., 2020).

We employ two primary loss functions to train the student model to ensure both global and fine-grained feature learning are effectively captured.

Global Loss (CLS Token): The global loss, applied to the [CLS] token, measures the difference between the probability distribution of the [CLS] token in the student model and the [CLS] token in the teacher model using KL divergence:

$$\mathcal{L}_{\text{CLS}} = \text{D}_{KL}\left(p_{\text{s}}^{[\text{CLS}]} \| p_{\text{t}}^{[\text{CLS}]}\right) \tag{6}$$

where $p_{\text{s}}^{[\text{CLS}]}$ is the predicted probability distribution of the [CLS] token from the student model, $p_{\text{t}}^{[\text{CLS}]}$ is the target probability distribution of the [CLS] token from the teacher model, D_{KL} represents the Kullback-Leibler divergence (Kullback & Leibler, 1951).

Fine-grained Loss (Patch): The patch loss also utilizes KL divergence, comparing the distribution of patch tokens in the student model to those in the teacher model:

$$\mathcal{L}_{\text{patch}} = \frac{1}{|\mathcal{M}|} \sum_{j \in \mathcal{M}} \text{D}_{KL}\left(p_{\text{s}}^{j} \| p_{\text{t}}^{j}\right) \tag{7}$$

where $\mathcal{M}$ is the set of indices corresponding to the masked patches, p_{s}^{j} is the predicted probability distribution of the j-th masked patch from the student model, p_{t}^{j} is the target probability distribution of the j-th masked patch from the teacher model, D_{KL} represents the Kullback-Leibler divergence.

3.2.2 PIXEL-LEVEL RECONSTRUCTION

As illustrated in Figure 1, following token removal, we employ an autoencoding method (g_ϕ-d_ϕ) (MAE (He et al., 2022)) to reconstruct the original signal from a partially observed input. By processing only visible patches, the encoder reduces data handling and improves efficiency. The decoder reconstructs the original image using the encoded visible patches and mask tokens representing the masked areas. Since the decoder is smaller and used only during pre-training, most computational effort is focused on training a more powerful encoder.

This process is especially well-suited for fine-grained tasks because it pushes the encoder to extract detailed information from the limited visible patches, sharpening its ability to recognize intricate features. By concentrating on these visible patches, the encoder becomes skilled at capturing subtle details that are critical in precision-demanding tasks like segmentation or detailed object recognition. The decoder's responsibility in reconstructing the masked portions of the image ensures that these fine details are preserved and accurately restored. The loss function ($\mathcal{L}_{\text{Rec}}$) employed during this training is the Mean Squared Error (MSE) between the reconstructed and original images, calculated solely on the masked patches:

$$\mathcal{L}_{\text{Rec}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|\hat{X}_i - X_i\|^2 \tag{8}$$

where $\mathcal{M}$ is the indices corresponding to the masked patches, $\hat{X}_i$ is the reconstructed patch for the i-th masked patch, X_i is the original patch for the i-th masked patch, $\| \cdot \|$ represents the Euclidean norm (or ℓ_2 norm).

3.2.3 TOKEN REMOVAL

Token removal in ViT refines the attention mechanism by eliminating less relevant tokens, reducing computational load, prioritizing important patches, and improving downstream performance. Existing methods include random token removal (Kakogeorgiou et al., 2022) and attentive removal (Yang et al., 2023). Our approach, tailored to the model's tasks—textual processing, patch comparison, and image reconstruction—outperforms these methods by effectively handling both fine-grained and coarse-grained details (see Section A.1 for more detail).

As shown in Figure 1, our token removal process begins by passing an image view through the teacher encoder (g_θ). The attention values generated by the encoder are used to mask the 50% of patches with the lowest values. For each patch p_t, the attention values are computed as:

$$AV_{p_t} = \frac{1}{N} \sum_{i=1}^{N} \text{Softmax} \left(\frac{Q_i \cdot K_i(p_t)}{\sqrt{d}} \right) \tag{9}$$

where N is the total number of attention heads across all layers, Q_i is the query of the [CLS] from the i-th attention head, $K_i(p_t)$ is the key of the patch at position p_t from the i-th attention head, d is the dimensionality of the query and key.

As shown in Figure 3 of the appendix, masking image patches by removing the lowest half of the attention values (AV) calculated by the teacher encoder results in the images displayed in the figure.

3.2.4 INTEGRATED LOSS FUNCTION

We introduce the formulation of a comprehensive loss function that combines the various loss components essential for effective multi-task learning. The integrated loss function is defined as:

$$\mathcal{L}_{\text{tot}} = \alpha_1 \times \mathcal{L}_{\text{CLS}} + \alpha_2 \times \mathcal{L}_{\text{Patch}} + \alpha_3 \times \mathcal{L}_{\text{Rec}} + \mathcal{L}_{\text{CLIP}}, \tag{10}$$

In this equation, the hyperparameters α_1, α_2, and α_3 determine the relative importance of each loss term. These weights are set to 1 in our experiments unless specified otherwise. The comprehensive loss function is designed to enable the model to concurrently learn across the four tasks introduced in previous sections, ensuring a balanced and integrated training process. In Section A.4 of the appendix, we explore the impact of different hyper-parameter choices, demonstrating how varying these weights can influence the overall model performance.

4 EXPERIMENTS

4.1 SETUP

Our experiments were conducted on a setup with four nodes, each equipped with four NVIDIA A100 GPUs (80GB). For details on computational efficiency, see Section A.2 in the appendix.

4.1.1 TRAINING DATA AND AUGMENTATION STRATEGY

Our model is trained on a carefully curated 15 million image subset of the YFCC100M dataset (Thomee et al., 2016), which contains only English-language titles and descriptions. For each image, a valid caption—either a title or description—is randomly selected during training, following the methodology used in SLIP (Mu et al., 2022).

To enhance model robustness, we employ a data augmentation strategy similar to that of SLIP. Images are randomly resized and cropped to a size between 50% and 100% of their original dimensions. This augmentation is applied to the images in the online training branches, allowing the model to learn from varied perspectives of the same image. For the teacher branch, we use a larger randomly cropped sub-image compared to the online views, which allows us to accurately compute attention value.

4.1.2 ARCHITECTURE AND TRAINING SETTING

Figure 1 illustrates the architecture of the DetailCLIP framework, which is composed of six key components: two visual encoders, denoted as g_θ, and g_ϕ, a text encoder named e_γ, a vision decoder named d_ϕ, and two heads referred to as h_θ and h_ϕ.

To ensure a fair comparison with existing models, our visual encoders (g_θ and g_ϕ) are based on the Vision Transformer (ViT-B/16) (DOSOVITSKIY, 2020) architecture. This model features 12 layers, each with a width of 768 and 12 attention heads.

For the text encoder (e_γ), we adopted a 12-layer Transformer architecture with a width of 512 and 8 attention heads, following the design principles established by CLIP (Radford et al., 2021).

The decoder (d_ϕ) itself is composed of a series of transformer blocks like MAE He et al. (2022). The decoder handles the entire set of input patches, including both the attention-masked and retained patches. Positional embeddings are added to each token in the set to preserve the spatial information.

The projection heads (h_θ and h_ϕ) in our framework are implemented as 3-layer MLPs, featuring an L2-normalized bottleneck, similar to the approach used in DINO (Caron et al., 2021). The output dimension of the shared projection head is set to 8192 like iBOT paper (Zhou et al., 2021), for robust feature representation and alignment during training.

For our experiments, we employ the $AdamW$ optimizer (Loshchilov & Hutter, 2017) with a learning rate of $5e - 4$ and a weight decay of 0.5. The training is conducted with a substantial batch size of 4096.

4.2 EXPERIMENTAL ANALYSIS

Additional experiments and analyses are provided in the appendix due to page limitations: Section A.1 (attention-based token removal), Section A.2 (compute efficiency), Section A.3 (pseudocode), Section A.4 (ablation studies), and Section A.5 (visual comparison of model outputs).

4.2.1 DETAIL-ORIENTED VISUAL TASKS

Methods	Dataset	Epoch	Effective View	ADE20K		COCO	
				UperNet	Linear	AP^b	AP^m
Self-Supervised							
DeiT	IN-1K	300	720M	47.4	-	44.1	39.8
MAE	IN-1K	800	960M	46.5	34.3	46.2	39.1
DINO	IN-1K	800	9600M	46.8	34.5	47.4	40.1
iBOT	IN-1K	300	720M	47.3	34.7	48.4	42.1
AttMask	IN-1K	300	4320M	47.5	35.2	48.9	42.2
CLIP-Based Model							
CLIP	400M	-	-	46.4	34.2	43.6	39.5
SLIP	YFCC-15M	25	750M	46.6	36.1	44.0	40.3
MaskCLIP	YFCC-15M	25	750M	47.5	36.3	45.4	40.9
A-CLIP	YFCC-15M	25	750M	47.0	34.7	45.8	41.7
DetailCLIP	YFCC-15M	25	750M	**48.1**	**37.3**	**48.9**	**42.5**
DetailCLIP	YFCC-15M	50	1500M	**48.8**	**39.3**	**50.1**	**43.3**

Table 1: Performance comparison on segmentation (ADE20K) and object detection (MS COCO) tasks using the ViT-B architecture. Segmentation results are based on UperNet and linear decoders, while object detection uses bounding box Average Precision. "Effective View" indicates the model's training exposure, calculated as views per image × epochs × dataset size.

We evaluated DetailCLIP on detail-oriented visual tasks to test its ability to capture fine-grained details in complex visual environments.

Semantic Segmentation on ADE20K: We evaluated DetailCLIP on the ADE20K dataset (Zhou et al., 2017) using both UperNet (Xiao et al., 2018) and linear decoders. UperNet assessed the model's ability to capture detailed object boundaries, while the linear decoder tested performance under minimal architectural complexity. Input images had a resolution of 512×512 pixels, and

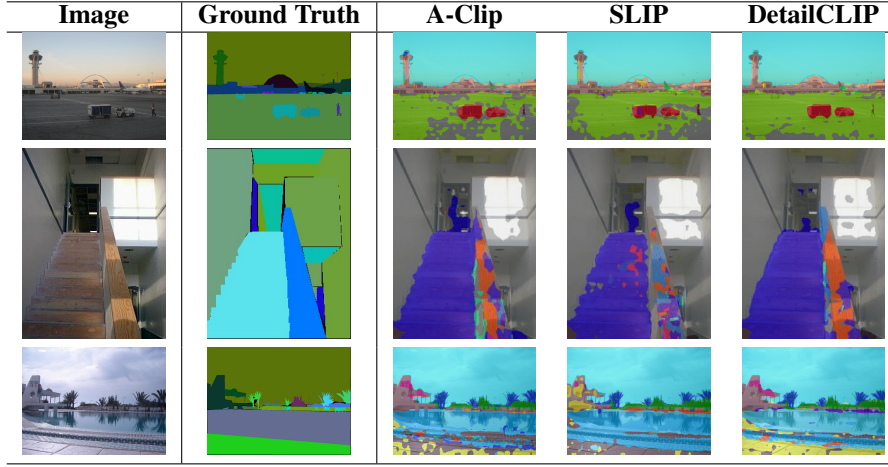| Image	Ground Truth	A-Clip	SLIP	DetailCLIP

Figure 2: Visual comparison of segmentation results using the Linear decoder across different models. The comparison includes SLIP, Attentive Mask CLIP, and DetailCLIP at 25 epochs.

training was conducted for $160k$ iterations. Performance was measured using the mean Intersection over Union (mIoU) metric.

Object Detection and Instance Segmentation on COCO: We evaluated DetailCLIP using the Cascade Mask R-CNN model (Cai & Vasconcelos, 2019; He et al., 2017), following iBOT, to generate bounding boxes and instance masks on the COCO dataset (Lin et al., 2014). Performance was measured using Average Precision for bounding boxes (AP^b) and masks (AP^m).

Evaluating Task Outcomes: Table 1 shows DetailCLIP's strong performance across detail-oriented visual tasks. Using the ViT-B architecture, DetailCLIP achieved a $mIoU$ of 48.8 on ADE20K with UperNet, surpassing MaskCLIP by 1.3 points. In the linear decoder setting, DetailCLIP's $mIoU$ of 39.3 outperformed SLIP by 3.2 points. In object detection on MS COCO, DetailCLIP achieved 48.9 AP^b for bounding boxes, surpassing ACLIP by 3.1 points, and 42.5 AP^m for masks, 0.8 points higher than the best alternative. These improvements highlight DetailCLIP's effectiveness in capturing both high-level semantics and fine-grained details.

Visual Comparative Analysis: Figure 2 shows the result of our model compared to the other baselines. Additionally, we invite you to refer to Section A.5 of the appendix, where we cover segmentation and object detection tasks.

4.2.2 IMAGE CLASSIFICATION

Zero-shot on Text-Image Retrieval: We also present the zero-shot text-image retrieval on 3 benchmark datasets: Flickr30K (Young et al., 2014), MS-COCO (Lin et al., 2014), and ImageNet-1K (Deng et al., 2009). Our findings indicate that using plain text, without any added prefixes or suffixes, consistently yields strong performance across all models evaluated. Table 2 comprehensively evaluates different models on text-image retrieval tasks, explicitly focusing on Flickr30K and MS-COCO datasets, also a zero-shot evaluation on ImageNet-1K.

At 25 epochs, DetailCLIP demonstrates the highest performance across most metrics, achieving the best results in Flickr I2T (62.8), T2I (42.2), COCO I2T (38.3), and zero-shot on IN-1K (43.9), closely matching A-CLIP. For 50 epochs, DetailCLIP maintains its superior performance, particularly in the COCO dataset, where it ties in with A-CLIP for the best I2T score (39.8) and surpasses A-CLIP in T2I with a score of 24.9. The results suggest that DetailCLIP is particularly strong in tasks that require detailed matching between text and images. DetailCLIP outperforms other models in both datasets' I2T and T2I retrieval tasks, indicating its effectiveness in text-image association.

Additionally, in Section A.4 of the appendix, we also provide the effect of different hyperparameter settings on ImageNet-1K's zero-shot scores.

Zero-Shot Classification Performance Evaluation on Diverse Benchmarks: We evaluated DetailCLIP's zero-shot classification performance on 13 tasks, following SLIP's (Mu et al., 2022) evaluation protocols with identical prompts for fair comparison. The results in the table 3 demonstrate that DetailCLIP consistently outperforms its competitors. Specifically, DetailCLIP achieved the highest performance in 12 out of 26 evaluation scenarios and secured the second-best position in 5 additional cases. DetailCLIP also achieved the highest average scores in both the 25-epoch (37.3) and 50-epoch (37.6) settings.

Methods	E	Flickr30K		COCO		IN-1K
		I2T	T2I	I2T	T2I	0-Shot
CLIP	25	51.4	32.6	27.9	17.6	37.6
SLIP	25	57.2	41.2	33.6	21.9	42.8
MaskCLIP	25	60.0	38.8	34.1	21.2	42.7
A-CLIP	25	62.7	42.1	38.0	**23.2**	**43.9**
DetailCLIP	25	**62.8**	**42.2**	**38.3**	22.9	**43.9**
CLIP	50	53.9	35.8	30.2	19.2	39.4
SLIP	50	60.6	41.1	33.2	22.3	44.1
A-CLIP	50	**66.7**	43.2	**39.8**	24.4	**46.3**
DetailCLIP	50	65.9	**44.7**	39.8	**24.9**	46.2

Table 2: Comparison of model performance on text-image retrieval for Flickr30K, MS COCO, and zero-shot ImageNet-1K (IN-1K). Metrics for image-to-text (I2T) and text-to-image (T2I) retrieval are shown after 25 and 50 epochs (E), based on results from A-CLIP Yang et al. (2023).

Epochs	Methods	SUN397	Aircraft	DTD	Pets	Flowers	GTSRB	MNIST	STL-10	KITTI	Country211	PCAM	CLEVR	SST2	Average
	CLIP	51.1	5.4	21.2	28.5	53.3	6.1	8.4	90.5	35.1	10.5	53.5	10.8	50.7	32.7
	SLIP	53.4	5.7	26.1	31.1	56.6	**14.5**	9.8	94.4	34.0	11.6	55.4	**17.5**	**51.1**	35.5
25	MaskCLIP	54.0	8.2	25.5	**36.8**	53.6	10.1	11.2	93.9	30.5	12.5	51.2	12.9	50.0	34.6
	A-CLIP	**57.0**	7.6	26.0	32.0	57.7	13.1	9.8	**95.4**	35.2	**13.5**	51.6	14.1	49.9	35.6
	DetailCLIP	54.9	**8.7**	**30.1**	35.4	**60.7**	9.3	**11.4**	94.1	**42.2**	11.9	**62.0**	13.4	50.2	**37.3**
	CLIP	49.0	6.3	23.5	27.2	56.2	7.4	**12.5**	92.1	33.6	10.9	50.8	14.0	50.1	33.4
50	SLIP	54.8	9.0	**29.8**	31.9	57.7	9.0	9.8	95.6	**35.1**	12.7	54.4	13.8	49.9	35.7
	A-CLIP	**58.7**	10.2	27.7	40.5	61.0	**9.4**	11.3	95.5	23.3	**14.4**	63.7	**19.6**	**52.3**	37.5
	DetailCLIP	57.5	**11.0**	29.8	**40.8**	**64.1**	7.3	10.8	**95.8**	30.9	13.3	63.9	12.7	50.1	**37.6**

Table 3: Zero-shot performance evaluation across multiple classification benchmarks. Epochs indicate the total number of training epochs. The results were collected in the A-CLIP paper Yang et al. (2023).

5 CONCLUSION

In this paper, we introduce DetailCLIP, a self-improving vision-language framework designed to overcome the limitations of traditional CLIP-based models in both fine- and coarse-grained tasks. While CLIP excels at aligning global image-text representations, it often misses the fine-grained details needed for tasks like segmentation and object detection. To address this, DetailCLIP incorporates self-curated learning objectives—patch-level self-distillation, pixel-level reconstruction, and attention-based token filtering—to iteratively refine both high-level semantics and detailed features. Experiments show that DetailCLIP generalizes well across tasks, surpassing state-of-the-art models on detail-focused benchmarks like ADE20K and MS COCO, while also delivering strong performance on coarse-grained tasks such as zero-shot classification.

ACKNOWLEDGMENTS

REFERENCES

Amani Almalki and Longin Jan Latecki. Self-supervised learning with masked autoencoders for teeth segmentation from intra-oral 3d scans. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7820–7830, 2024.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

Ido Ben-Shaul, Ravid Shwartz-Ziv, Tomer Galanti, Shai Dekel, and Yann LeCun. Reverse engineering self-supervised learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1483–1498, 2019.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11165–11174, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020b.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9640–9649, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, 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.

Alexey DOSOVITSKIY. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Bahar Farahani and Amin Karimi Monsefi. Smart and collaborative industrial iot: A federated learning and data space approach. *Digital Communications and Networks*, 9(2):436–447, 2023.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Wenbin He, Suphanut Jamonnak, Liang Gou, and Liu Ren. Clip-s4: Language-guided self-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11207–11216, 2023.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Siyu Jiao, Yunchao Wei, Yaowei Wang, Yao Zhao, and Humphrey Shi. Learning mask-aware clip representations for zero-shot segmentation. *Advances in Neural Information Processing Systems*, 36:35631–35653, 2023.

Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. In *European Conference on Computer Vision*, pp. 300–318. Springer, 2022.

Amin Karimi Monsefi, Pouya Shiri, Ahmad Mohammadshirazi, Nastaran Karimi Monsefi, Ron Davies, Sobhan Moosavi, and Rajiv Ramnath. Crashformer: A multimodal architecture to predict the risk of crash. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Advances in Urban-AI*, pp. 42–51, 2023.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. *arXiv preprint arXiv:2408.04883*, 2024.

Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23390–23400, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 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.

Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. 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.

Ran Liu, Ellen L Zippi, Hadi Pouransari, Chris Sandino, Jingping Nie, Hanlin Goh, Erdrin Azemi, and Ali Moin. Frequency-aware masked autoencoders for multimodal pretraining on biosignals. *arXiv preprint arXiv:2309.05927*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, pp. 23033–23044. PMLR, 2023.

Amin Karimi Monsefi, Payam Karisani, Mengxi Zhou, Stacey Choi, Nathan Doble, Heng Ji, Srinivasan Parthasarathy, and Rajiv Ramnath. Masked logonet: Fast and accurate 3d image analysis for medical domain. *arXiv preprint arXiv:2402.06190*, 2024.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*, pp. 529–544. Springer, 2022.

Pouyan Navard and Alper Yilmaz. A probabilistic-based drift correction module for visual inertial slams. *arXiv preprint arXiv:2404.10140*, 2024.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Shehan Perera, Pouyan Navard, and Alper Yilmaz. Segformer3d: an efficient transformer for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4981–4988, 2024.

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.

Qing Su, Anton Netchaev, Hai Li, and Shihao Ji. Flsl: Feature-level self-supervised learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11686–11695, 2022.

Yixuan Wei, Yue Cao, Zheng Zhang, Houwen Peng, Zhuliang Yao, Zhenda Xie, Han Hu, and Baining Guo. iclip: Bridging image classification and contrastive language-image pre-training for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2776–2786, 2023.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 418–434, 2018.

Jiahao Xie, Wei Li, Xiaohang Zhan, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. Masked frequency modeling for self-supervised visual pre-training. In *The Eleventh International Conference on Learning Representations*, 2023.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9653–9663, 2022.

Yifan Yang, Weiquan Huang, Yixuan Wei, Houwen Peng, Xinyang Jiang, Huiqiang Jiang, Fangyun Wei, Yin Wang, Han Hu, Lili Qiu, et al. Attentive mask clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2771–2781, 2023.

Kun Yi, Yixiao Ge, Xiaotong Li, Shusheng Yang, Dian Li, Jianping Wu, Ying Shan, and Xiaohu Qie. Masked image modeling with denoising contrast. In *The Eleventh International Conference on Learning Representations*, 2023.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36, 2024.

Behzad Zakeri, Amin Karimi Monsefi, Sanaz Samsam, and Bahareh Karimi Monsefi. Weakly supervised learning technique for solving partial differential equations; case study of 1-d reaction-diffusion equation. In *High-Performance Computing and Big Data Analysis: Second International Congress, TopHPC 2019, Tehran, Iran, April 23–25, 2019, Revised Selected Papers 2*, pp. 367–377. Springer, 2019.

Tianyi Zheng, Bo Li, Shuang Wu, Ben Wan, Guodong Mu, Shice Liu, Shouhong Ding, and Jia Wang. Mfae: Masked frequency autoencoders for domain generalization face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 2024.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

Jinghao Zhou, Li Dong, Zhe Gan, Lijuan Wang, and Furu Wei. Non-contrastive learning meets language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11028–11038, 2023a.

Mengxi Zhou, Yue Zhang, Amin Karimi Monsefi, Stacey S Choi, Nathan Doble, Srinivasan Parthasarathy, and Rajiv Ramnath. Reducing manual labeling requirements and improved retinal ganglion cell identification in 3d ao-oct volumes using semi-supervised learning. *Biomedical Optics Express*, 15(8):4540–4556, 2024.

Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11175–11185, 2023b.

Methods	Training Time	GPU Memory
CLIP	1.00x	14G
SLIP	2.67x	30G
MaskCLIP	1.56x	16G
A-CLIP	1.16x	14G
DetailCLIP	1.16x	24G

Table 4: Training time and GPU memory usage of various models, measured consistently on the same device. Training time is reported relative to the baseline CLIP (1.00x), and GPU memory is shown in gigabytes (GB) Yang et al. (2023).

A APPENDIX

A.1 ATTENTION TOKEN REMOVAL

a person crafting pottery on a potter's wheel, installed in a lowered section of the floor.

Two dogs sitting on the grass; one is a golden retriever, and the other is a basset hound.

A black bird with a red beak perched on a tree branch against a blue sky.

A large stupa with a golden spire in a mountainous landscape under a blue sky.
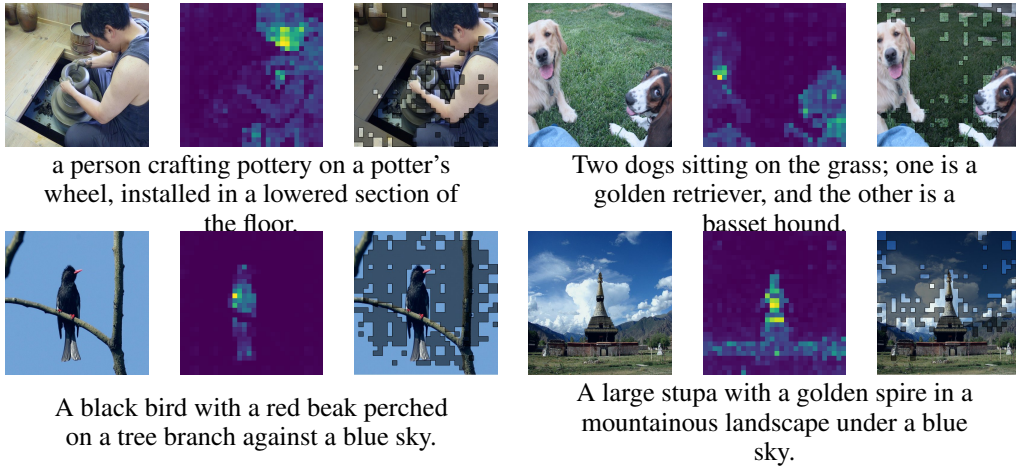
Figure 3: Illustration of the token removal mechanism, which retains semantically significant tokens to focus on critical input aspects while discarding unnecessary details.

In Figure 3, the mechanism of token removal is illustrated. The process selectively eliminates less relevant tokens, retaining only those with strong semantic significance. This approach enhances the model's focus on the most important aspects of the input, ensuring that critical information is preserved while unnecessary details are discarded. The current settings of the model use a 50% mask ratio, preserving only the semantically relevant details of the image. For example, in the image with the two dogs, the grass background is removed as it is less important to the overall image. Similarly, in the image of the black bird, the tree and branches are ignored while preserving the bird, as it is central to the image. In the DetailCLIP model, the attention scores are derived from the teacher network and are used to guide the student network during training.

A.2 COMPUTATIONAL EFFICIENCY

Our model demonstrates an effective balance between performance and efficiency, as highlighted in Table 4 and Table 2. It requires only 24G of GPU memory, significantly less than SLIP (30G), while achieving a training time of 1.16x—closely aligned with CLIP (1.00x) and notably faster than SLIP (2.67x) and MaskCLIP (1.56x). Our resource usage and training speed make DetailCLIP highly suitable for applications with limited GPU resources, while offering strong performance on downstream tasks.

A.3 PSEUDOCODE

Algorithm 1 presents the pseudocode for the DetailCLIP algorithm, which integrates CLIP, self-distillation, and pixel-level loss. The steps include updating the momentum encoder using EMA,

generating teacher embeddings for both views (without gradients), and creating attention-based masks with a 50% mask ratio. Student embeddings are then obtained after applying the masks and projected using CLIP projections alongside text embeddings. The model produces outputs for student distillation and reconstruction, followed by computing classification, patch-level, and reconstruction losses to calculate the total loss.

Algorithm 1 Pytorch-like code for DetailCLIP

```python
# u, v: view 1 and view 2 of the image
# s_visual, t_visual: student and teacher image encoders
# t_dist_head, s_dist_head: teacher and student distillation heads
# encode_image, encode_text: students CLIP image and text projections

# reconstr_head: pixel-level reconstruction head
def forward(u, v, text, momentum):
    # Update momentum encoder using EMA
    update_momentum_encoder(momentum)

    # Teacher embeddings for both views (no gradient)
    with no_grad():
        t_embed_u, attn_u = t_visual(u)
        t_embed_v, attn_v = t_visual(v)
        t_dist_out = t_dist_head(concat(t_embed_u, t_embed_v))

    # Generate masks from attention using 50% mask ratio
    mask_u, mask_v = get_att_mask(attn_u), get_att_mask(attn_v)

    # Get student embeddings after applying masks
    s_embed_u = s_visual(u, mask=mask_u)
    s_embed_v = s_visual(v, mask=mask_v)

    # Project student embeddings using CLIP projections
    img_emb_u = encode_image(s_embed_u)
    img_emb_v = encode_image(s_embed_v)

    # Get text embedding
    text_emb = encode_text(tokenize(text))

    # Student distillation output
    s_dist_out = s_dist_head(concat(s_embed_u, s_embed_v))

    # Reconstruction
    s_dec_u = decoder(s_embed_u)
    s_dec_v = decoder(s_embed_v)
    u_rec = h_reconst(s_dec_u)
    v_rec = h_reconst(s_dec_v)

    # Compute losses
    clip_loss = compute_clip_loss(img_emb_u, img_emb_v, text_emb)
    distill_loss = compute_distill_loss(s_dist_out, t_dist_out)
    recon_loss = compute_recon_loss(u_rec, v_rec, mask_u, mask_v)

    # Total loss
    total_loss = clip_loss + distill_loss + recon_loss
    return total_loss
```

A.4 ABLATION STUDY ON LOSS WEIGHT

In this ablation study, we investigate the impact of different weight configurations on the comprehensive loss function $\mathcal{L}_{tot}$ introduced in Eq. 10. We aim to identify the optimal values for the hyperparameters α_1, α_2, and α_3 that effectively balance the various losses.

	$\alpha_1 = 1$ $\alpha_2 = 1$ $\alpha_3 = 1$	$\alpha_1 = 1$ $\alpha_2 = 1$ $\alpha_3 = 0$	$\alpha_1 = 1$ $\alpha_2 = 1$ $\alpha_3 = 2$	$\alpha_1 = .5$ $\alpha_2 = .5$ $\alpha_3 = .5$	$\alpha_1 = 0$ $\alpha_2 = 0$ $\alpha_3 = 1$
Acc	**43.9**	42.9	43.2	43.3	42.6

Table 5: Impact of varying α values on the zero-shot accuracy of ImageNet-1k after 25 epochs of training with the DetailCLIP.

Table 5 presents the results of this exploration, showing how different combinations of these weights influence the zero-shot accuracy on the ImageNet-1k dataset after 25 epochs of training. The accuracy values reflect the model's performance under varying degrees of emphasis on each loss term.

Analyzing the results, we find that the baseline configuration, where all weights are set to 1, achieves the highest accuracy of 43.9%, suggesting that equal emphasis on all loss components provides a robust performance baseline. When α_3 is reduced to 0, thereby ignoring the reconstruction loss, the accuracy slightly decreases to 42.9%, indicating that while this loss component is not critical, it still positively influences overall model performance. Doubling α_3 to 2 results in a marginal accuracy decrease to 43.2%, implying that an overemphasis on the reconstruction loss can detract slightly from the model's effectiveness in other tasks. Lowering the weights of α_1 and α_2 to 0.5 yields an accuracy of 43.3%, demonstrating that moderate weight reductions do not significantly impact performance. However, eliminating these ($\alpha_1 = 0$ and $\alpha_2 = 0$) reduces accuracy further to 42.6%, underscoring their essential role in the model's learning process.

A.5 VISUAL ANALYSIS

This section presents a comprehensive comparative analysis of DetailCLIP's performance against several baseline models, primarily focusing on segmentation and object detection tasks. Through a series of detailed visualizations, we provide side-by-side comparisons that emphasize the superior capability of DetailCLIP in capturing and reconstructing fine-grained details, which is crucial for achieving high segmentation accuracy. Specifically, Figure 4 showcases the semantic segmentation results using the UPerNet decoder, while Figure 5 highlights the results with a linear decoder. These figures emphasize the notable improvements achieved by DetailCLIP in handling complex images, particularly when it comes to accurately delineating boundaries and preserving detailed features, areas where baseline models often struggle. The segmentation comparisons are conducted using the ADE20K dataset.

In addition to segmentation, Figures 6 and 7 present the object detection results using the MS COCO dataset, further demonstrating the effectiveness of DetailCLIP across multiple vision tasks. By visualizing these outputs, we showcase DetailCLIP's enhanced performance and shed light on the limitations of other models, such as SLIP and A-CLIP. These comparisons underscore the significant advancements made by DetailCLIP in tackling detail-oriented tasks, solidifying its robustness and versatility in a wide array of visual challenges.

Overall, these visual and quantitative evaluations illustrate the significant advancements made by DetailCLIP in addressing detail-oriented challenges across both segmentation and object detection. The side-by-side comparisons solidify DetailCLIP's robustness and adaptability, making it a more effective and versatile solution for a wide range of fine-grained vision tasks.
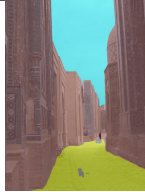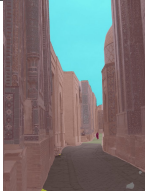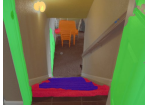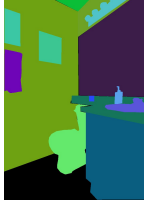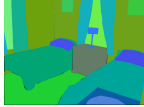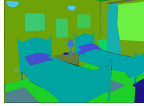
Figure 4: Visual comparison of segmentation results using the UPerNet decoder across different models. The comparison includes SLIP, A-CLIP, DetailCLIP at 25 epochs, and DetailCLIP at 50 epochs.
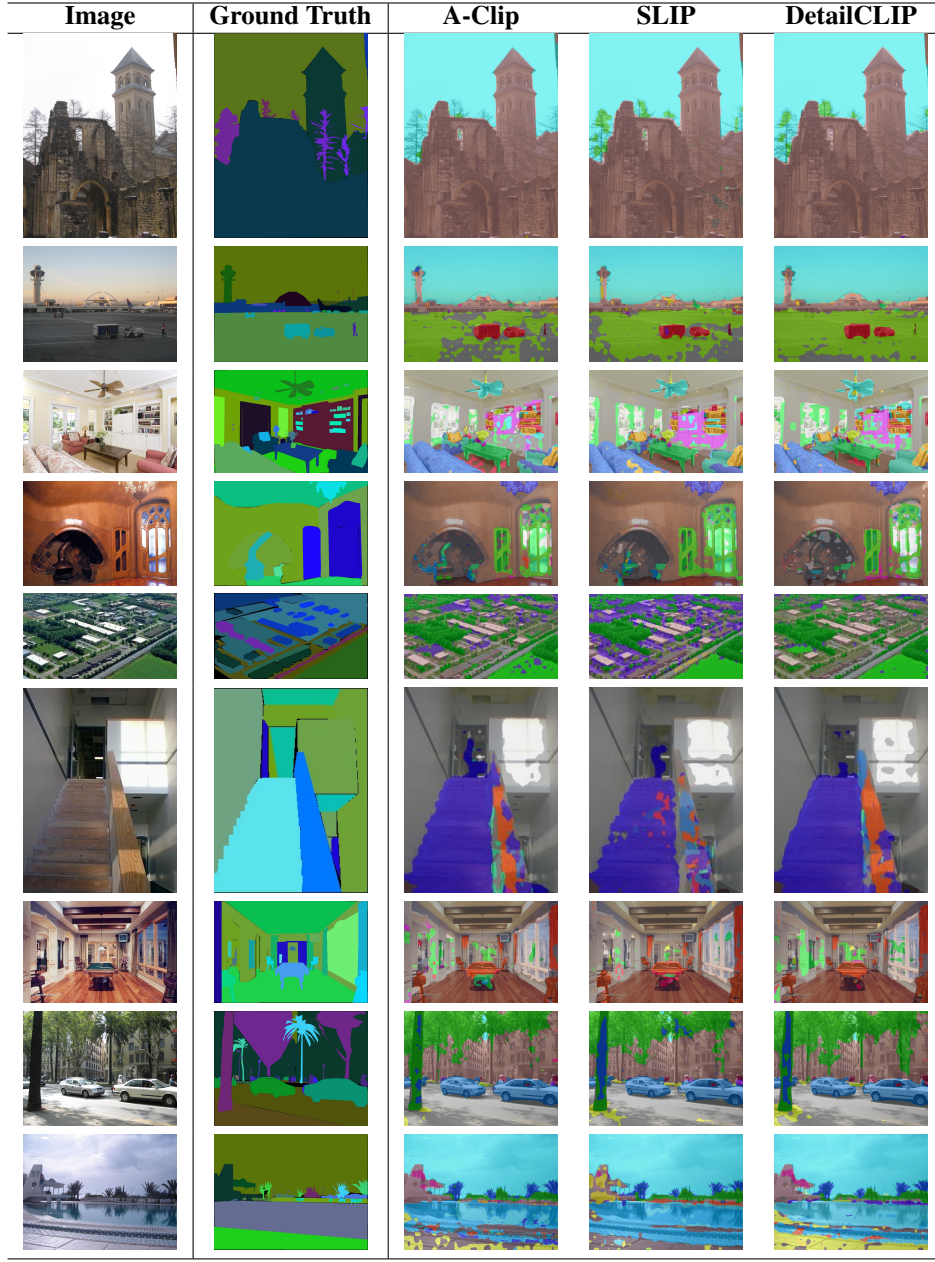
Figure 5: Visual comparison of segmentation results using the Linear decoder across different models. The comparison includes SLIP, A-CLIP, and DetailCLIP at 25 epochs.
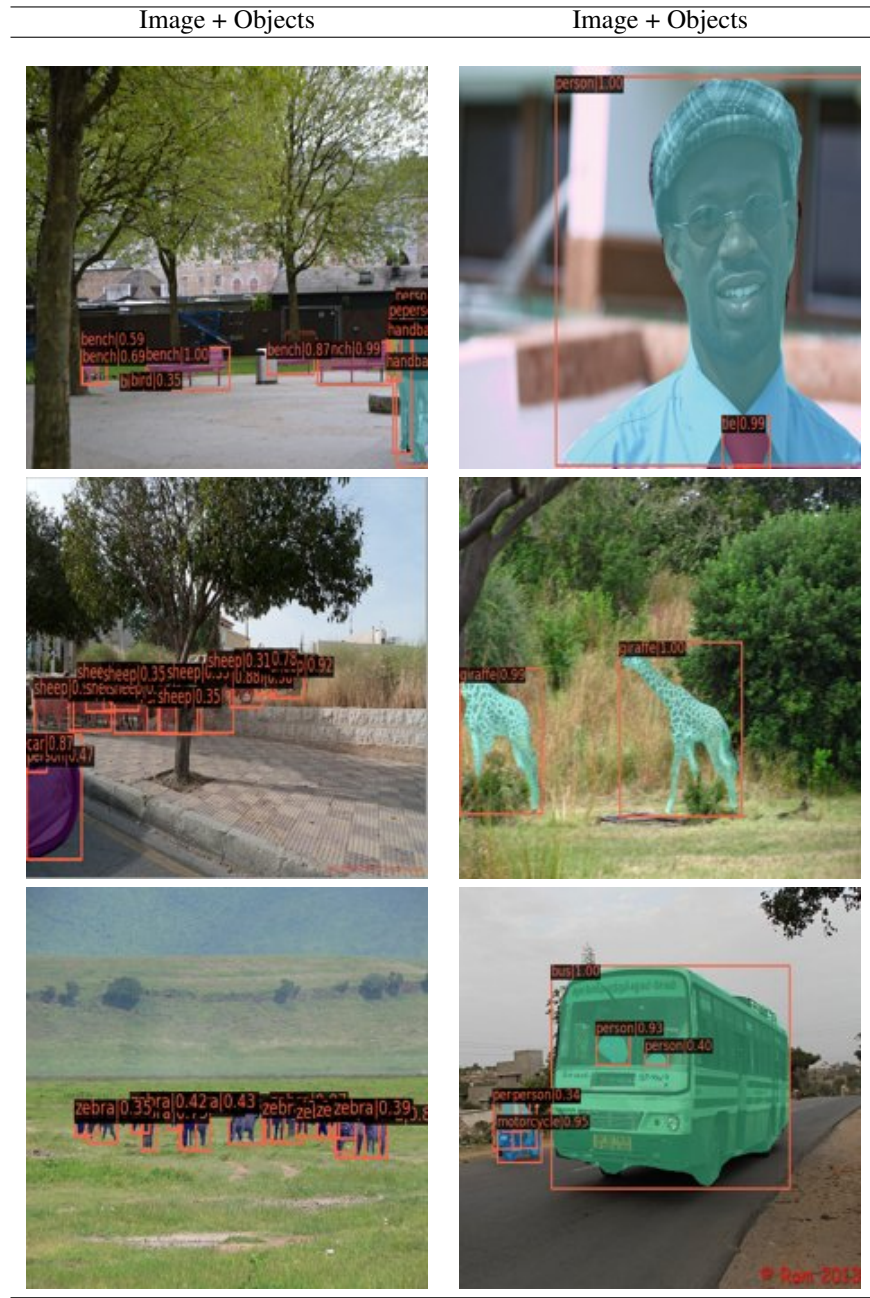
Image + Objects	Image + Objects

Figure 6: Object detection results using DetailCLIP after 50 training epochs. The image highlights DetailCLIP's ability to detect objects accurately, showcasing its precision and effectiveness in complex scenes

Image + Objects	Image + Objects

Figure 7: Object detection results using DetailCLIP after 50 training epochs. The image highlights DetailCLIP's ability to detect objects accurately, showcasing its precision and effectiveness in complex scenes