Appendix

In the supplementary materials, we provide more implementation details and ablations on the hyper-parameters of our CFM-ViT design. We also present more visualizations of the feature reconstruction in our masked image-text pretraining and discuss limitation of CFM-ViT.

A. Implementation Details

Table 1 and 2 summarize the hyperparameters used in our masked image-text pretraining and open-vocabulary detection finetuning, respectively.

B. More Ablations

Our default setting uses the fixed 2D sinusoidal PE in the ViT backbone. In Table 3a, we compare this with the trainable PE and observed no benefits. Note that masked feature reconstruction is not added in this experiment.

In Table 3b, we ablate the number of the reconstruction decoder blocks and observe two decoder blocks work the best. Table 3c ablates the loss coefficient between the contrastive and reconstruction losses. We set $L_{\text{con}} : L_{\text{rec}} = 1 : 2$ as our default setting.

C. Visualizations: Feature Reconstruction

Fig. 1 provides qualitative examples of the feature reconstruction task in our masked image-text pretraining (Sec. 3.2). We test the pretrained model on the Flickr image-text paired dataset. Our reconstruction branch takes a heavily masked image (middle image per example) and predicts the masked features in the joint image-text embedding space. We visualize the similarity map between the reconstructed image features and a query text embedding (right image per example). We observe that the learned reconstructions are semantically plausible with respect to the queried image-text pairs.

D. Limitations

CFM-ViT utilizes the knowledge in pretrained Vision Language Models. The resulting detector model weights will reflect the data biases. In this paper, we mainly demonstrate CFM-ViT’s capabilities in comparison with existing open-vocabulary detection works.
### Table 3: Ablation study on LVIS open-vocabulary detection benchmark. ViT-L/16 backbone and contrastive batch size 4k are used unless otherwise noted. Note that masked feature reconstruction is not used in subtable (a). Our best setting is marked by gray.

<table>
<thead>
<tr>
<th>PE Type</th>
<th>AP_r</th>
<th>AP</th>
<th># dec. blocks</th>
<th>AP_r</th>
<th>AP</th>
<th>$L_{con}:L_{rec}$</th>
<th>AP_r</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>fixed sinusoidal</td>
<td>27.4</td>
<td>30.4</td>
<td>1</td>
<td>30.0</td>
<td>33.7</td>
<td>1:1</td>
<td>30.2</td>
<td>33.3</td>
</tr>
<tr>
<td>trainable</td>
<td>27.2</td>
<td>30.3</td>
<td>2</td>
<td>30.7</td>
<td>34.0</td>
<td>1:2</td>
<td>30.4</td>
<td>34.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>29.9</td>
<td>34.0</td>
<td>1:5</td>
<td>30.3</td>
<td>34.0</td>
</tr>
</tbody>
</table>

(a) Fixed sinusoidal PE vs trainable PE.

(b) Number of recon. decoder blocks.

(c) Loss weights between $L_{con}$ and $L_{rec}$.

---

![Feature reconstruction visualization](image1.png)

“A heavy machine lifting up a worker”

“A woman vending street food”

“A dog catching a frisbee”

“A young person is kayaking in blue water”

Figure 1: Feature reconstruction visualization. For each example, we visualize the (left) original image, (middle) masked image, and (right) the similarity map between the reconstructed features and the text query embedding (bottom). We observe that our CFM-ViT is able to predict whole-image semantics from the heavily masked image.

---

**E. Dataset License**

- COCO [3]: Creative Commons Attribution 4.0 License
- LVIS [2]: CC BY 4.0 + COCO license
- COCO Captions (retrieval) [1]: CC BY
- Flickr30k (retrieval) [4]: Custom (research-only, non-commercial)
- Objects365 [5]: Custom (research-only, non-commercial)

**References**


