Reconstructing CLIP for Open-Vocabulary Dense Perception

Supplementary Material

6. The Necessity of Spatial Aggregation for A_p

We present the multi-layer expansion of the spatial fusion alignment loss. Let the spatially fused student representation be

$$\tilde{v}_r^{(S)} = \sum_{l=1}^{L}\sum_{j=1}^{N} S_{rj} P^{(l)} v_j^{(l)}. \tag{16}$$

Decompose the fusion into on-target ($j = r$) and off-target ($j \in \mathcal{U}_r$) parts:

$$\begin{aligned} T_{\text{on}} &\triangleq \sum_{l=1}^{L} (S_{rr} P^{(l)} - M^{(l)}) v_r^{(l)}, \\ T_{\text{off}} &\triangleq \sum_{l=1}^{L} \sum_{j\in\mathcal{U}_r} S_{rj} P^{(l)} v_j^{(l)}. \end{aligned} \tag{17}$$

The expected MSE alignment loss then has the following compact expansion:

$$\begin{aligned} L^{(S)} = \mathbb{E}\big\|\tilde{v}_r^{(S)} - c_r\big\|^2 &= \mathbb{E}\|T_{\text{on}}\|^2 + \mathbb{E}\|T_{\text{off}}\|^2 + \mathbb{E}\|\varepsilon_r\|^2 \\ &+ 2\,\mathbb{E}\langle T_{\text{on}}, T_{\text{off}}\rangle - 2\,\mathbb{E}\langle T_{\text{on}}, \varepsilon_r\rangle - 2\,\mathbb{E}\langle T_{\text{off}}, \varepsilon_r\rangle. \end{aligned} \tag{18}$$

Under mild assumptions commonly used for deriving interpretable bounds, the cross-correlations between on-target and off-target features are small, and the residual ε_r can be treated as teacher noise that is approximately independent of the feature components. Under these conditions, the cross terms in Eq. (18) contribute only weakly compared with the main quadratic terms. In contrast, the off-target quadratic component $\mathbb{E}\|T_{\text{off}}\|^2$ remains substantial because it accumulates contributions from all unrelated spatial locations. This term reflects the intrinsic effect of spatial aggregation and cannot be eliminated. In that regime the dominant additional contribution relative to the head-only modeling design is the off-target quadratic term, which can be approximated as

$$\mathbb{E}\|T_{\text{off}}\|^2 \approx \sum_{l=1}^{L}\sum_{j\in\mathcal{U}_r} S_{rj}^2\, \mathbb{E}\|P^{(l)} v_j^{(l)}\|^2, \tag{19}$$

recovering the interference term used in the main text and justifying the claim that spatial aggregation introduces an unavoidable positive penalty when attention assigns weight to unrelated patches.

7. Experimental Settings

7.1. Implementation Details

Within the distillation framework, every input image is subdivided into a grid configuration of $m \times n$ cells. Here, m and n are integers randomly sampled from 1 to 6. The cropped image corresponding to each grid region is individually resized to 224×224 pixels and processed by the frozen teacher model to obtain its associated CLS token, which is subsequently used in the semantics distillation. Unless stated otherwise, our experiments are conducted using EVA-CLIP. For the spatial coherence distillation, we account for the mismatch in patch granularity between CLIP and DINOv2 (e.g., patch sizes of 16 and 14, respectively). To ensure that both models produce the same number of visual tokens, we adjust their input resolutions accordingly. For instance, we use an input size of 1024 for the student VLM and 896 for DINOv2, allowing both to generate exactly 4096 image tokens, thereby enabling token-wise correlation alignment during distillation. The feature coherence distillation loss follows prior work [35] and aims to align the spatial correlation structures between the student and teacher representations. Specifically, given the L2-normalized dense features of the student model $\mathcal{X}^s_{dense} \in \mathbb{R}^{N\times D}$ and the corresponding features $\mathcal{X}^t_{dense} \in \mathbb{R}^{N\times D}$ from the teacher DINOv2, we construct patch-wise correlation matrices to capture the pairwise similarities between spatial locations. The coherence loss is defined as:

$$\begin{aligned} \mathbf{G}^s &= \mathcal{X}^s_{dense}(\mathcal{X}^s_{dense})^T, \\ \mathbf{G}^t &= \mathcal{X}^t_{dense}(\mathcal{X}^t_{dense})^T, \\ \mathcal{L}_{coherence} &= \frac{1}{N^2}\left\|\mathbf{G}^s - \mathbf{G}^t\right\|_2. \end{aligned} \tag{20}$$

DenseRC maintains high computational efficiency without incurring significant additional overhead. For 1024^2 inputs with ViT-B/16, the method requires 636 GFLOPs (a reduction of 57 GFLOPs compared to the 693 GFLOPs in DeCLIP) and introduces only 0.009M parameters on top of DeCLIP's 86M.

7.2. Open-Vocabulary Detection

For the open-vocabulary detection, we follow the common distillation protocol adopted in prior works [35, 37]. Images are resized to 1024×1024, and dense features are distilled over region proposals generated by an RPN pretrained on COCO train2017. We adopt the two-stage F-ViT framework as the downstream detector, keeping its architecture and training pipeline unchanged to ensure a fair comparison.

We evaluate DenseRC on four benchmarks: OV-COCO [21], OV-LVIS [10], COCO, and Objects365 [30]. OV-COCO contains 48 base and 17 novel categories. Only the base categories (107,761 images) are used for training.

Table 8. Detailed comparison with state-of-the-art methods for open-vocabulary object detection. Caption supervision denotes that the model is trained with additional image–text pairs, while CLIP supervision indicates semantic transfer from the original CLIP model. FineCLIP leverages CC2.5M to generate region–text pairs as supervision.

(a) OV-COCO benchmark [21]

Method	Supervision	Backbone	AP_{50}^{Novel}	AP_{50}^{Base}	AP_{50}
ViLD [9]	CLIP	RN50	27.6	59.5	51.2
Detic [53]	Caption	RN50	27.8	51.1	45.0
OV-DETR† [47]	CLIP	RN50	29.4	61.0	52.7
ProxyDet [11]	Caption	RN50	30.4	52.6	46.8
RegionCLIP [51]	Caption	RN50	31.4	57.1	50.4
RTGen [2]	Caption	RN50	33.6	51.7	46.9
BARON-KD [36]	CLIP	RN50	34.0	60.4	53.5
CLIM [38]	CLIP	RN50	36.9	-	-
SAS-Det [50]	CLIP	RN50	37.4	58.5	53.0
RegionCLIP [51]	Captions	RN50x4	39.3	61.6	55.7
CORA† [39]	CLIP	RN50x4	41.7	44.5	43.8
OV-DQUO† [34]	CLIP	RN50x4	45.6	-	-
RO-ViT [16]	CLIP	ViT-L/16	33.0	-	47.7
CFM-ViT [15]	CLIP	ViT-L/16	34.1	-	46.0
F-ViT [37]	CLIP	ViT-B/16	37.6	54.9	50.4
BIND [49]	CLIP	ViT-L/16	41.5	58.3	54.8
F-ViT [37]	CLIP	ViT-L/14	44.3	64.1	59.0
F-ViT+FineCLIP [14]	CC2.5M	ViT-B/16	29.8	45.9	41.7
F-ViT+FineCLIP [14]	CC2.5M	ViT-L/14	40.0	57.2	52.7
F-ViT+CLIPSelf [37]	CLIP	ViT-B/16	37.6	54.9	50.4
F-ViT+CLIPSelf [37]	CLIP	ViT-L/14	44.3	64.1	59.0
F-ViT+DeCLIP [35]	CLIP	ViT-B/16	41.1	57.8	53.5
F-ViT+DeCLIP [35]	CLIP	ViT-L/14	46.2	65.2	60.3
F-ViT+DenseRC	CLIP	ViT-B/16	45.6	57.9	54.6
F-ViT+DenseRC	CLIP	ViT-L/14	54.8	65.4	62.6

(b) OV-LVIS benchmark [10]

Method	Supervision	Backbone	mAP_r	mAP_c	mAP_f	mAP
ViLD [9]	CLIP	RN50	16.6	24.6	30.3	25.5
OV-DETR† [47]	CLIP	RN50	17.4	25.0	32.5	26.6
BARON-KD [36]	CLIP	RN50	22.6	27.6	29.8	27.6
RegionCLIP [51]	Caption	RN50x4	22.0	32.1	36.9	32.3
CORA+† [39]	Caption	RN50x4	28.1	-	-	-
SAS-Det [50]	CLIP	RN50x4	29.1	32.4	36.8	33.5
CLIM [38]	CLIP	RN50x64	32.3	-	-	-
F-VLM [17]	CLIP	RN50x64	32.8	-	-	34.9
F-ViT [37]	CLIP	ViT-B/16	25.3	21.8	29.1	25.2
RTGen [2]	Caption	Swin-B	30.2	39.9	41.3	38.8
BIND [49]	CLIP	ViT-L/16	32.5	33.4	35.3	33.2
Detic [53]	Caption	Swin-B	33.8	-	-	47.0
CFM-ViT [15]	CLIP	ViT-L/14	33.9	-	-	36.6
RO-ViT [16]	CLIP	ViT-H/16	34.1	-	-	35.1
F-ViT [37]	CLIP	ViT-L/14	34.9	34.6	35.6	35.1
ProxyDet [11]	Caption	Swin-B	36.7	-	-	41.5
CoDet [25]	Caption	ViT-L/14	37.0	46.3	46.3	44.7
OV-DQUO† [34]	CLIP	ViT-L/14	39.3	-	-	-
F-ViT+FineCLIP [14]	CC2.5M	ViT-B/16	10.4	8.0	10.9	9.5
F-ViT+FineCLIP [14]	CC2.5M	ViT-L/14	20.2	19.6	23.1	20.9
F-ViT+CLIPSelf [37]	CLIP	ViT-B/16	25.3	21.8	29.1	25.2
F-ViT+CLIPSelf [37]	CLIP	ViT-L/14	34.9	34.6	35.6	35.1
F-ViT+DeCLIP [35]	CLIP	ViT-B/16	26.8	22.4	29.8	26.0
F-ViT+DeCLIP [35]	CLIP	ViT-L/14	37.2	35.2	36.5	36.0
F-ViT+DenseRC	CLIP	ViT-B/16	29.0	22.6	30.6	26.8
F-ViT+DenseRC	CLIP	ViT-L/14	39.6	35.4	37.7	37.0

Table 9. Zero-shot cross-dataset transfer evaluation of the LVIS-trained detector on COCO and Objects365.

Method	COCO [21]			Objects365 [30]					
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Supervised Baseline [9]	46.5	67.6	50.9	25.6	38.6	28.0	-	-	-
ViLD [9]	36.6	55.6	39.6	11.8	18.0	12.6	-	-	-
DetPro [6]	34.9	53.8	37.4	12.1	18.8	12.9	4.5	11.5	18.6
BARON [36]	36.2	55.7	39.1	13.6	21.0	14.5	5.0	13.1	20.7
F-VLM [17]	37.9	59.6	41.2	16.2	25.3	17.5	-	-	-
CoDet [25]	39.1	57.0	42.3	14.2	20.5	15.3	-	-	-
RO-ViT [16]	-	-	-	17.7	27.4	19.1	-	-	-
F-ViT[37]+FineCLIP [14]	33.6	52.7	36.1	12.1	19.8	12.6	2.3	13.3	29.3
F-ViT[37]+CLIPSelf [37]	40.5	63.8	44.3	19.5	31.3	20.7	9.7	23.2	35.5
F-ViT[37]+DeCLIP[35]	41.0	64.6	44.8	20.0	32.2	21.2	10.0	24.4	36.7
F-ViT[37]+DenseRC (Ours)	43.4	66.6	47.6	20.8	32.4	22.6	10.1	24.8	37.5

The validation split includes 4,836 images covering both base and novel classes. Following standard practice, we report the mean Average Precision (mAP) at an Intersection over Union (IoU) threshold of 0.5. OV-LVIS spans 1,203 categories in total. Its training set comprises 461 common and 405 frequent categories (100,170 images), while the validation set contains 19,809 images covering the full spectrum (frequent, common, and rare). We report the mAP at IoU thresholds ranging from 0.5 to 0.95.

During downstream finetuning, we simply replace the CLIP visual backbone in F-ViT with our DenseRC. All training procedures and hyperparameters strictly follow the official settings in [35]. Importantly, the HSG module remains frozen throughout finetuning to ensure that improvements come solely from our distilled representation. Comprehensive results on OV-COCO, OV-LVIS, and cross-dataset settings are summarized in Tab. 8 and Tab. 9. DenseRC consistently surpasses the previous approaches across all evaluation protocols, with particularly large gains on unseen (novel) categories, highlighting both the generalization capability and effectiveness of our dense representation learning framework.

7.3. Open-Vocabulary Semantic Segmentation

For the OVSS task, we adopt the distillation setup used in prior studies [35]. The dense representations are distilled from features pooled over regular grid partitions of the image.

For downstream training, we follow the standard OVSS pipeline and train on the COCO-Stuff [1] dataset, which offers 118k finely annotated images across 171 semantic categories. Evaluation is carried out on two popular benchmarks: ADE20K [52] and PASCAL-Context [7]. ADE20K provides 2k validation images. Following convention, we report performance on both the 150-class subset (A-150) and the full 847-class set (A-847). PASCAL-Context includes 5,104 validation images, and we evaluate on the full 459-category set (PC-459) as well as the commonly used 59-class subset (PC-59). Performance is measured using mean Intersection-over-Union (mIoU), computed by averaging per-class IoU scores, in line with previous OVSS literature [4, 13, 40].

For the segmentation model, we adopt CAT-Seg [4] as the downstream framework. Its architecture, training recipe,

and hyperparameter configurations are strictly kept intact to ensure a fair comparison. The only modification is substituting its default visual encoder with our DenseRC. Similar to our detection setup, the HSG module is frozen during finetuning, allowing us to attribute downstream improvements solely to the distilled dense representations.

8. Additional Experiments

Integration with MLLMs. To evaluate DenseRC's impact on fine-grained perception in multimodal large language models (MLLMs), we integrate DenseRC as the vision encoder into the LLaVA framework. The results are summarized in Tab. 14. Specifically, when equipping LLaVA-1.5-7B [22] (which originally utilizes the OpenAI CLIP ViT-L/14 [28] as its visual backbone) with DenseRC under identical experimental settings, our method consistently improves performance. Remarkably, it even surpasses LLaVA1.5-13B on GQA and POPE benchmarks, demonstrating enhanced fine-grained perception and reduced hallucination in MLLMs.

Results on SigLIP v1/v2. We further integrate DenseRC into SigLIP v1 [48] and v2 [33] to evaluate its robustness across different VLMs. The results are summarized in Tab. 10. As shown, DenseRC consistently brings substantial improvements in dense perception performance. Notably, it significantly boosts the recognition accuracy of stuff masks from 0.04%/0.03% to over 54%, highlighting its effectiveness. These results clearly demonstrate the generalization ability of DenseRC across different VLMs.

Table 10. Results on SigLIP v1/v2. DenseRC consistently improves dense perception across different VLMs.

Method	Boxes	Thing Masks	Stuff Masks
SigLIP v1-B/16	43.2	42.6	0.04
SigLIP v1-B/16 + DenseRC	**77.9**	**78.9**	**55.6**
SigLIP v2-B/16	46.9	46.7	0.03
SigLIP v2-B/16 + DenseRC	**81.5**	**82.1**	**54.8**

Visualization. Fig. 6 presents qualitative comparisons of open-vocabulary detection on the OV-COCO benchmark. In these visualizations, detections for novel classes are highlighted in red, while base class predictions are shown in blue. Notably, CLIPSelf fails to detect the novel category *elephant* in the first example and misclassifies the novel *cake* as *bowl* in the third. In contrast, our DenseRC yields precise detections for both base and novel categories. These results underscore our model's superior capability in localizing and recognizing unseen objects.

Fig. 7 illustrates open-vocabulary semantic segmentation results. The model is trained exclusively on COCO-Stuff and evaluated on ADE20K in a zero-shot setting, demonstrating its robust capacity to generalize dense semantics across diverse datasets and categories. Notably, CLIP-Self yields imprecise masks for novel categories such as *painting* and *table* in the first example. In contrast, our DenseRC produces more accurate segmentation masks, correctly identifying these challenging novel categories.

9. Ablation Study

Core Components. We conduct a systematic ablation study to quantify the contributions of $\mathcal{L}_{\text{semantics}}$ and $\mathcal{L}_{\text{coherence}}$. Additionally, to verify the robustness of our approach across different visual foundation models (VFMs), we replace DINOv2 with DINOv3 [31] for comparative experiments.

As shown in Tab. 11, each component provides substantial performance gains. Incorporating only $\mathcal{L}_{\text{semantics}}$ increases the object mask recognition accuracy from 18.2% to 75.3%, corresponding to a 56.1% relative improvement, while using only $\mathcal{L}_{\text{coherence}}$ results in a 50.8% relative gain. When both constraints are applied jointly, the model achieves further improvements, demonstrating their complementary effect.

Regarding adaptability to different VFMs, our method maintains strong dense feature representation capabilities with DINOv3, which employs larger patch sizes that tend to reduce spatial granularity. Notably, the mean patch feature similarity in DINOv3's final layer is considerably higher than in DINOv2, indicating over-smoothing that could degrade fine-grained semantic information. To mitigate this, we extract and distill coherence from the penultimate layer of DINOv3 to better preserve distinct patch semantics of CLIP.

Table 11. Ablation study on core components. We report Top1 mean accuracy.

	Boxes	Thing Masks	Stuff Masks
EVA-CLIP	18.2	20.6	18.4
+ $\mathcal{L}_{\text{semantics}}$	75.3	76.8	50.5
+ $\mathcal{L}_{\text{coherence}}$ (DINOv2)	65.9	71.4	51.9
$\mathcal{L}_{\text{semantics}}$ + $\mathcal{L}_{\text{coherence}}$ (DINOv2)	76.7	78.1	55.8
$\mathcal{L}_{\text{semantics}}$ + $\mathcal{L}_{\text{coherence}}$ (DINOv3)	76.3	78.1	55.8

Training Data. We evaluate the robustness of DenseRC across different training data distributions. As shown in Tab. 15, our method consistently yields performance gains in dense prediction tasks, even when trained on limited data (e.g., 1,464 samples from Pascal VOC). This demonstrates its effectiveness across varying data scales.

W_H. We adopt a shared W_H across layers in the HSG module. Importantly, sharing W_H does not yield identical head gating weights across layers, since the weights are computed from diverse head input features. As layers in HSG jointly construct $\mathcal{X}_{\text{dense}}$, sharing W_H enables effective cross-layer modeling while adding few parameters, reducing overfitting with slight gains over per-layer W_H, as

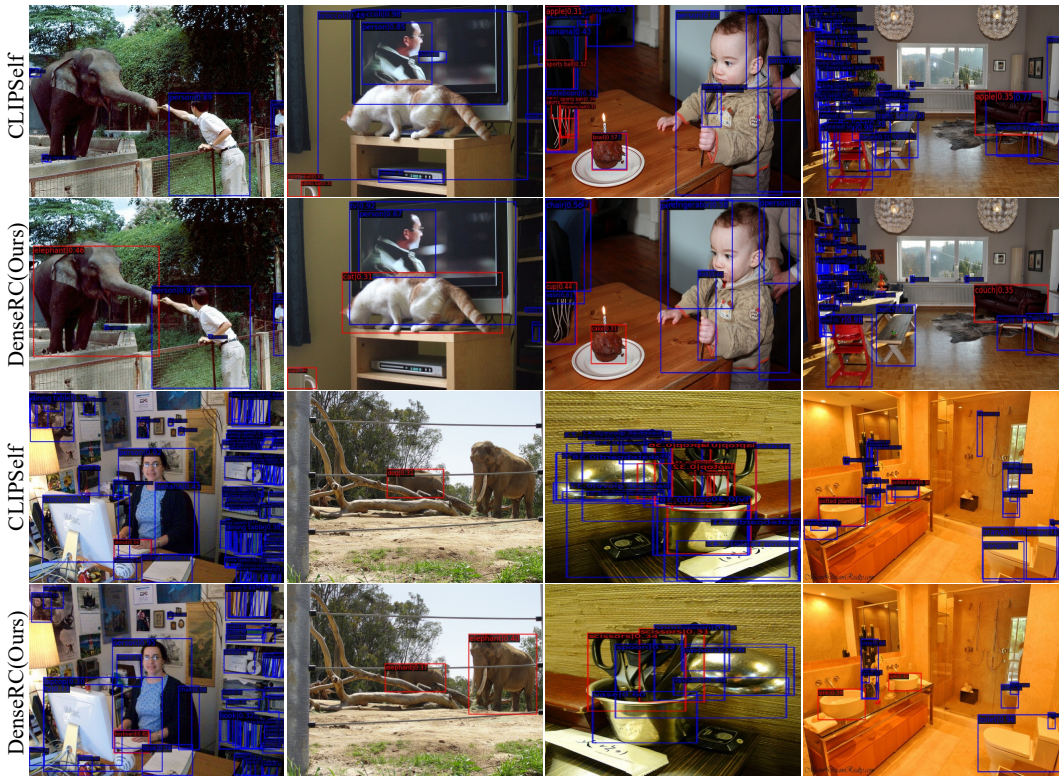

Figure 6. Visualization of open-vocabulary detection results on OV-COCO. Red bounding boxes denote novel categories, while blue ones correspond to base categories.

Table 12. Effect of sharing W_H in the HSG module. Sharing W_H across layers yields better performance with fewer parameters compared to using layer-specific W_H, demonstrating improved generalization and cross-layer modeling.

W_H	Boxes	Thing Masks	Stuff Masks
per-layer	76.6	77.9	55.6
shared	**76.7**	**78.1**	**55.8**

Table 13. Ablation study on final MLP.

Final MLP in teacher	Final MLP in student	A-847	PC-459	A-150	PC-59
✓	✓	15.9	22.6	37.6	61.4
✓	✗	15.9	22.7	37.6	61.3
✗	✗	15.6	22.8	37.2	61.2

shown in Tab. 12.

MLP. Recent OVSS studies [18, 19] observe that the final MLP (FFN block) in CLIP degrades dense representations. To further investigate the role of the MLP under our self-distillation framework, we conduct ablations by selectively enabling or disabling the final MLP in the teacher or student models. The results are summarized in Tab. 13. As shown, including the FFN in the student model only causes marginal changes in performance, indicating that our framework is insensitive to the final MLP. Interestingly, removing the final MLP in the frozen teacher model also does not lead to significant performance drops, suggesting that the FFN contributes little to the image representations, consistent with prior observations [8]. Based on these findings, we retain the original FFN in the teacher CLIP model by default, while removing it from the student model to better align with dense downstream methods.

Table 14. Integration of DenseRC into MLLMs. Replacing the original CLIP-based visual encoder in LLaVA1.5-7B with DenseRC consistently improves performance across benchmarks.

MLLM	visual encoder	image size	GQA	Text VQA	POPE rand	POPE pop	POPE adv	MME
LLaVA1.5 -7B	CLIP-L	336	62.0	58.2	87.3	86.1	84.2	1510.7
	+DenseRC	336	**63.4**	**58.9**	**88.3**	**87.1**	**85.7**	**1525.0**
LLava1.5-13B	CLIP-L	336	63.3	61.3	87.1	86.2	84.5	1531.3

Table 15. Ablation study on training data. We report Top1 mean accuracy.

Method	Backbone	Training data	Image samples	Boxes	Thing Masks	Stuff Masks
EVA-CLIP	ViT-B/16	✗	✗	18.2	20.6	18.4
+DenseRC	ViT-B/16	PASCAL VOC	1464	70.0	71.6	52.5
+DenseRC	ViT-B/16	ADE20K	20210	74.8	76.8	54.9

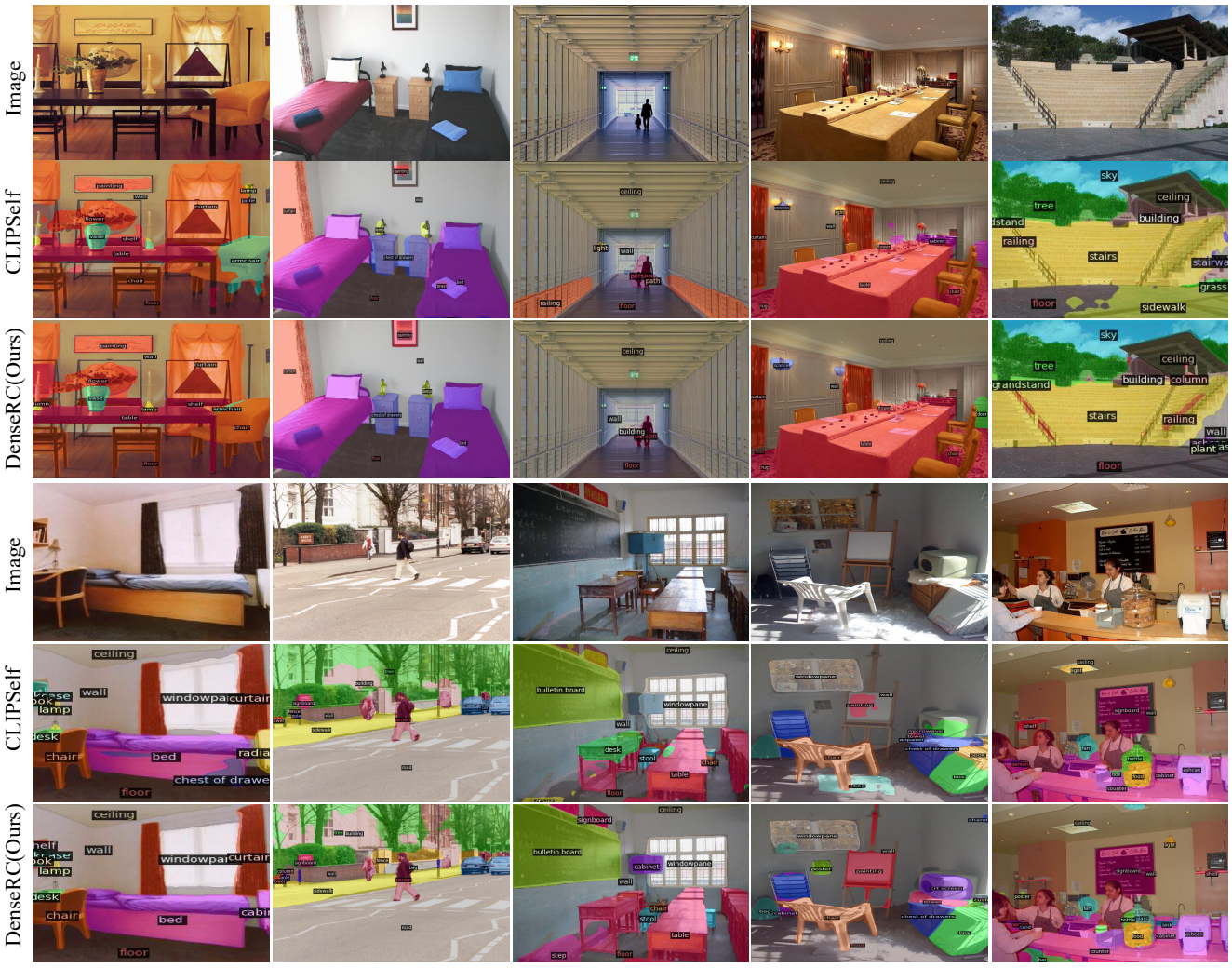

Figure 7. Visualization of open-vocabulary semantic segmentation results on ADE20K (A-150).