Talking to DINO: Bridging Self-Supervised Vision Backbones with Language for Open-Vocabulary Segmentation

Luca Barsellotti*[1], Lorenzo Bianchi*[2,3], Nicola Messina[2], Fabio Carrara[2], Marcella Cornia[1], Lorenzo Baraldi[1], Fabrizio Falchi[2], Rita Cucchiara[1]
[1]University of Modena and Reggio Emilia, Italy [2]ISTI-CNR, Italy [3]University of Pisa, Italy
[1]{name.surname}@unimore.it [2]{name.surname}@isti.cnr.it

Abstract

Open-Vocabulary Segmentation (OVS) aims at segmenting images from free-form textual concepts without predefined training classes. While existing vision-language models such as CLIP can generate segmentation masks by leveraging coarse spatial information from Vision Transformers, they face challenges in spatial localization due to their global alignment of image and text features. Conversely, self-supervised visual models like DINO excel in fine-grained visual encoding but lack integration with language. To bridge this gap, we present Talk2DINO, a novel hybrid approach that combines the spatial accuracy of DINOv2 with the language understanding of CLIP. Our approach aligns the textual embeddings of CLIP to the patch-level features of DINOv2 through a learned mapping function without the need to fine-tune the underlying backbones. At training time, we exploit the attention maps of DINOv2 to selectively align local visual patches with textual embeddings. We show that the powerful semantic and localization abilities of Talk2DINO can enhance the segmentation process, resulting in more natural and less noisy segmentations, and that our approach can also effectively distinguish foreground objects from the background. Experimental results demonstrate that Talk2DINO achieves state-of-the-art performance across several unsupervised OVS benchmarks. Source code and models are publicly available at: https://lorebianchi98.github.io/Talk2DINO/.

1. Introduction

Open-Vocabulary Segmentation (OVS) [59] is a fundamental task in Computer Vision that aims to partition an input image into a set of coherent regions based on concepts provided at inference time [6, 17, 27]. The set of concepts used to partition the image is usually provided in freeform natural language, which effectively unchains the methods from working with a specific fixed set of classes provided at training time. Solving this task requires a fine-grained understanding of the semantic interconnections between image pixels and the meaning conveyed by natural language.

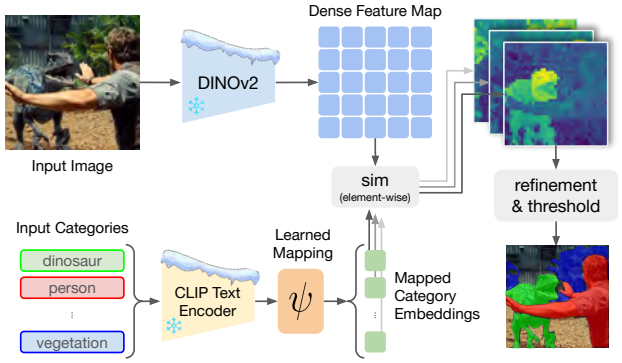

Figure 1. **Overview.** Our approach aligns the embedding spaces of CLIP and DINOv2 through a learned mapping function. This results in fine-grained visual encodings, which can be mapped to language to provide natural and less noisy semantic segmentation.

While previous works employed pixel-level annotations as a source of supervision [53, 58, 65], a recent trend in the field is to tackle this problem in an unsupervised manner [3, 9, 54] leveraging the correlations learned by state-of-the-art backbones. Contrastive embedding spaces like CLIP [38], in particular, have demonstrated good performance on tasks that demand a holistic understanding of vision and language modalities [28, 33, 60, 64], and have therefore been employed for unsupervised OVS [26, 47, 63]. Although CLIP-based backbones exhibit strong cross-modal capabilities, they are primarily trained to predict a global similarity score between text and images, which limits their spatial understanding and consequently affects tasks based on dense predictions. Recent efforts have tackled this limitation by introducing architectural modifications [18, 47, 63]. However, the spatial understanding constraints imposed by the training modality hinder the effectiveness of such backbones in OVS and highlight the potential benefits of exploring alternative models with enhanced perceptual capabilities.

*Equal contribution.

Self-supervised vision-only backbones like DINO and DINOv2 [8, 12, 37] have instead shown remarkable abilities in capturing fine-grained and localized spatial features without the reliance on annotated data. Specifically, the self-attention mechanism in such backbones generates attention maps that consistently pinpoint relevant regions within the image and has been widely leveraged for foreground object segmentation [42, 43, 48–50]. While this property makes them a powerful tool for tasks requiring fine-grained spatial understanding, the embedding space derived from visual self-supervised networks is not inherently aligned with textual concepts, making it incompatible with the OVS task.

To close the gap between vision-language and self-supervised embedding spaces, we propose Talk2DINO, a method that combines the spatial sensitivity of DINOv2 with the text-image alignment capabilities of CLIP, enabling a highly localized multimodal image understanding. Our approach, depicted in Fig. 1, learns a mapping function that translates the text embeddings of CLIP to interact with the patch-level embeddings of DINOv2 without fine-tuning the underlying backbones. Our alignment mechanism enhances text-image correspondence by exploiting the self-attention heads of DINOv2 to highlight diverse regions of the image. During training, we weight visual patch embeddings using the attention maps of DINOv2, dynamically selecting the head that best aligns with the provided caption. This embedding is then used to maximize similarity with the caption representation through contrastive learning. At inference time, we calculate the similarity of visual patches to each textual label, including a novel background cleaning procedure that weights class scores using attention maps.

Our approach demonstrates state-of-the-art performance in unsupervised OVS with minimal parameter learning. Our results show that a self-supervised vision-only encoder can generate embeddings with semantic properties akin to textual representations, opening up new pathways for addressing spatial understanding limitations in CLIP-like models [45]. To sum up, our main contributions are as follows:

- We propose Talk2DINO, the first model that provides language properties to DINOv2 by mapping the CLIP textual embeddings into the DINOv2 space through a non-linear warping function.
- Our proposed model employs a novel training schema that selects the most relevant visual self-attention head and does not need fine-tuning on the backbones.
- We showcase the capabilities of Talk2DINO on unsupervised OVS by devising a computationally efficient inference pipeline that also employs a novel approach based on DINOv2 self-attention to improve distinguishing foreground categories from the background.
- Experimentally, we show that Talk2DINO achieves state-of-the-art results in standard OVS benchmarks, demonstrating the effectiveness of the proposed approach.

2. Related Work

Vision-Language Pre-Training. In the last years, vision-language pre-training has gained increasing interest by learning multimodal representations that can be easily transferred to downstream tasks [16, 20]. The popular CLIP model [38] is trained on large-scale web-scraped data through a contrastive objective, which maximizes the similarity of the representations of corresponding image-text pairs while minimizing the similarity of the other pairs within a batch. This approach has demonstrated remarkable zero-shot classification and retrieval performance. However, learning a multimodal representation by matching global images and texts poses challenges in localizing regions with their corresponding text, showing limited performance in dense prediction tasks [4, 29, 35, 39, 60].

Open-Vocabulary Segmentation. In zero-shot segmentation, a segmentation model is trained on a set of *seen* classes and must generalize to *unseen* classes. The first attempts of OVS inherit this paradigm by training the model on a closed set of classes for which segmentation data is available and exploiting vision-language pre-training to extend their capabilities on an open set of classes through text [10, 19, 53, 58, 65]. The two-step approach represents the most popular framework in this research direction, which proposes class-agnostic regions, trained on segmentation masks, and provides them to CLIP to be aligned with textual classes [13, 17, 30, 56, 57]. However, this approach is affected by performance gaps between seen and unseen classes and presents a significant computational overhead.

On the contrary, another research direction investigates how to force the segmentation capabilities to emerge without relying on direct supervision from segmentation data. Cha *et al*. [9] identify this setting as unsupervised OVS and propose a unified evaluation protocol. Approaches in this line can be categorized into two main groups: (i) training-free methods that propose architectural adaptations to enable pre-trained models to produce localized multimodal features [5, 18, 25, 26, 44, 46, 47, 51, 63], and (ii) weakly-supervised methods that leverage large sets of image-text pairs with dedicated learning strategies that aim to improve the correspondence between regions and texts [9, 32, 39, 40, 54, 55]. Our model lies in the latter category since we exploit a weak supervision to learn how to bridge the CLIP and DINOv2 feature spaces.

Self-Supervised Backbones. Recent advances in self-supervised learning have led to models that showcase impressive matching and localization capabilities. In particular, the DINO family of models [8, 12, 37] employs Vision Transformers [14] trained with self-distillation, and has shown that patch-level features learned through self-supervision can yield semantic information. Also, a strong relationship has been observed between self-attention acti-

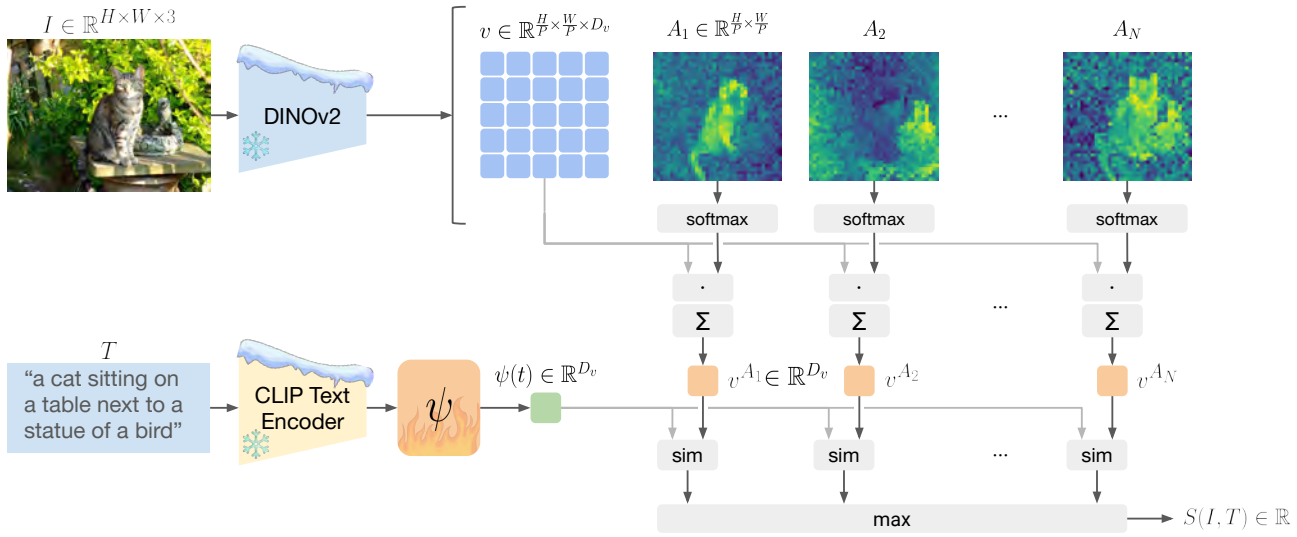

Figure 2. **Overview of the training methodology of Talk2DINO.** We learn a projection $\psi(\cdot)$ that maps the CLIP textual embeddings to the visual embedding space of DINOv2. Given the dense feature map and the attention maps extracted from DINOv2, we generate N visual embeddings by computing a weighted average of the feature map with each attention map. We then compute the similarity between each visual embedding and the projected text embedding, and use the maximum similarity as the global alignment score.

vations and foreground regions of the image. Hence, researchers have recently focused on the usage of DINO for the unsupervised segmentation task [42, 43, 48–50].

While the aforementioned models are purely visual, recent works have also focused on connecting self-supervised feature spaces with textual representations for OVS. ReCo [41], OVDiff [23], FOSSIL [2], and FreeDA [3] are training-free approaches that build prototypes in the visual space according to pre-defined textual categories. CLIP-DINOiser [52] demonstrates that CLIP can be fine-tuned with the supervision of DINO to retain improved localization capabilities. LaVG [22] employs DINO to propose class-agnostic regions and computes the average embedding from CLIP for each region. ProxyCLIP [26] proposes a proxy attention module to integrate DINO features with *values* from the last attention layer of CLIP, thus employing the two visual backbones at prediction time. Our method is closely related to this research field since we aim to combine the multimodal understanding capabilities of CLIP with the localization properties of DINOv2. However, in contrast to these methods, we propose to directly map the textual representations from the textual encoder of CLIP to the DINOv2 space, and demonstrate that our approach sets a new state-of-the-art without relying on multiple visual backbones or on external sources of knowledge.

3. Proposed Method

3.1. Preliminaries

Task Definition. Open-vocabulary segmentation aims to segment objects of interest defined through natural language at inference time. Let $I \in \mathbb{R}^{H \times W \times 3}$ be an image and $v(I) \in \mathbb{R}^{\frac{H}{P} \times \frac{W}{P} \times D_v}$ its dense feature map extracted by a Transformer-based visual backbone with input patch size P and dimensionality of embedding space D_v. Let $\{T_j\}_{j=1,...,M}$ be a set of arbitrary textual categories and $t(T_j) \in \mathbb{R}^{D_t}$ their embeddings extracted by a pre-trained textual backbone. To simplify the notation, in the following, we will refer to $v(I)$ as v and to $t(T_j)$ as t_j. Assuming a multimodal setting in which $D_t = D_v$, we could define the similarity map $\mathcal{S}(I, T_j) \in \mathbb{R}^{\frac{H}{P} \times \frac{W}{P}}$ for the image I and category T_j as the cosine similarity between t_j and each spatial entry of v. Formally, the similarity map is defined as

$$\mathcal{S}(I, T_j)_{[h,w]} = \frac{v_{[h,w]} \cdot t_j^{\intercal}}{||v_{[h,w]}|| \, ||t_j||}, \tag{1}$$

where $\cdot_{[h,w]}$ represents indexing over spatial axes. The full resolution similarity map $\hat{\mathcal{S}}(I, T) \in \mathbb{R}^{H \times W}$ is recovered by upsampling $\mathcal{S}(I, T)$ (*e.g.*, via bilinear interpolation). Segmentation masks $\mathcal{M}(I, T_1, ..., T_M)$ are then derived by assigning pixels to the category with the highest similarity score, *i.e.*,

$$\mathcal{M}(I, T_1, ..., T_M)_{[h,w]} = \underset{j=1,...,M}{\operatorname{argmax}} \, \hat{\mathcal{S}}(I, T_j)_{[h,w]}. \tag{2}$$

In order for Eq. 1, and therefore the segmentation from Eq. 2, to work correctly, not only the two v and t spaces should share the same dimensionality, but they should also be constructed so that they also share the same semantics.

CLIP and DINO Duality. Existing vision-language models trained on image-text pairs, such as CLIP [38], can naturally fit the formulation mentioned above, as they provide dense visual and textual embeddings in the same space.

However, while CLIP can correctly align global features coming from texts and images (*i.e.*, through the similarities corresponding to CLS tokens), it lacks a precise alignment between the textual feature t and spatial patches v.

Conversely, purely visual self-supervised backbones like DINOv2 [37] have shown remarkable semantic and local consistency of spatial embeddings, enabling agnostic image segmentation [42, 43, 50]. These abilities occur naturally in the last attention layer of DINOv2, where the attention maps computed between the CLS token and the spatial tokens align with relevant objects within the image (see Fig. 2). Despite the remarkable results observed on image-only tasks, DINOv2 lacks a solid bridge with natural language, making it impossible to directly compute the similarities with the text features, as expressed in Eq. 1.

While DINOv2 and CLIP embedding spaces are traditionally thought as being uncorrelated spaces, we show that the CLIP textual embedding space can be projected into the DINOv2 space through a learnable nonlinear warping.

3.2. Augmenting DINO with Semantics

Warping CLIP Embedding Space. We learn a projection $\psi : \mathbb{R}^{D_t} \rightarrow \mathbb{R}^{D_v}$ to map textual embeddings t into the space of the visual patch embeddings v of DINOv2, leveraging weak supervision from image-text pairs. We build the projection ψ applied to textual features by composing two affine transformations with a hyperbolic tangent activation, which provides nonlinear warping. Formally,

$$\psi(t) = \boldsymbol{W}_b^\intercal \left(\tanh(\boldsymbol{W}_a^\intercal t + b_a)\right) + b_b, \tag{3}$$

where $\boldsymbol{W}_a \in \mathbb{R}^{D_t \times D_v}$ and $\boldsymbol{W}_b \in \mathbb{R}^{D_v \times D_v}$ are learnable projection matrices and b_* are learnable bias vectors.

Mapping DINO to the Warped CLIP Space. To learn the nonlinear projection ψ, we exploit the intrinsic segmentation capability of DINOv2 to identify the precise spatial subsets of v to which $\psi(t)$ should be aligned with.

Specifically, we first extract the N attention maps $A_i \in \mathbb{R}^{\frac{H}{P} \times \frac{W}{P}}$ (one for each of the $i = 1, ..., N$ heads) which DINOv2 computes between the CLS vector and its patch features from the last layer. One of the key features of DINOv2 is that each A_i highlights different semantic regions within the image. For each attention map A_i, we compute a visual embedding $v^{A_i} \in \mathbb{R}^{D_v}$ as a weighted average of the dense feature map v, emphasizing the spatial areas that A_i highlights. We then calculate the cosine similarity between each v^{A_i} and the projected text embedding $\psi(t)$, resulting in N similarity scores. Formally, the cosine similarity score between a head and the text embedding is defined as

$$\text{sim}(v^{A_i}, t) = \frac{v^{A_i} \cdot \psi(t)^\intercal}{||v^{A_i}||\,||\psi(t)||}, \tag{4}$$

$$\text{with } v^{A_i} = \sum_{h,w} v_{[h,w]} \text{softmax}(A_i)_{[h,w]}. \tag{5}$$

To obtain the most relevant score for alignment, we apply a selection function over the similarity scores obtained for different heads. In particular, we choose the maximum similarity $\max_{i=1,...,N} \text{sim}(v^{A_i}, t)$ score across all heads, therefore promoting a robust alignment between textual and visual representations that adapts to the most salient visual features corresponding to the text query.

Training Procedure. To optimize the alignment between text and visual embeddings, we employ the InfoNCE loss, which leverages similarity scores across a batch of image-text pairs. For each pair (I_i, T_i), we compute similarity scores between the projected text embedding $\psi(t_i)$ and the maximally-activated visual embedding $\tilde{v}_i$, where $\tilde{v}_i$ is the visual embedding derived from the most relevant attention head for the corresponding text t_i, *i.e.*,

$$\tilde{v}_i = v_i^{A_j} \mid j = \underset{k=1,...,N}{\text{argmax}}\, \text{sim}(v_i^{A_k}, t_i). \tag{6}$$

Treating the true image-text pair as the positive instance and the remaining pairs within the batch as negatives, this contrastive approach drives the model to increase similarity for matching pairs and decrease it for non-matching pairs. Formally, the InfoNCE loss $\mathcal{L}_{\text{InfoNCE}}$ for a batch of B image-text pairs is defined as

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{2B} \sum_{i=1}^{B} \log \frac{\exp(\text{sim}(\tilde{v}_i, t_i))}{\sum_{j=1}^{B} \exp(\text{sim}(\tilde{v}_j, t_i))} - \frac{1}{2B} \sum_{i=1}^{B} \log \frac{\exp(\text{sim}(\tilde{v}_i, t_i))}{\sum_{j=1}^{B} \exp(\text{sim}(\tilde{v}_i, t_j))}.$$

This formulation effectively strengthens alignment by maximizing the similarity for true pairs and minimizing it for mismatched pairs across the batch.

Inference. The projection learned during the training procedure that warps the CLIP embedding space into the DINOv2 space enables the textual embeddings to be directly comparable with the dense feature embeddings from DINOv2. Hence, at inference time, given an image $I \in \mathbb{R}^{H \times W \times 3}$ and a set of textual arbitrary categories $\{T_j\}_{j=1,...,M}$, we can obtain the segmentation masks as defined in Eq. 2 by considering the projected text embeddings $\psi(t_j)$ in the similarity map computation from Eq. 1.

3.3. Identifying Background Regions

An additional challenge that OVS approaches need to face, especially when tasked with benchmarks like Pascal VOC [15] and COCO Objects [7], is that of identifying "background" regions, *i.e.* regions that do not belong to the set of categories considered in the benchmark. The standard approach consists in applying a threshold on the similarity or probability score to identify where the model is not

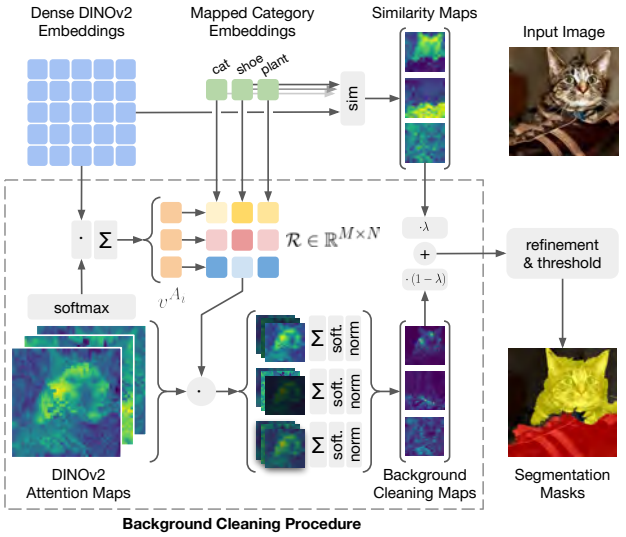

Figure 3. **Inference procedure.** At the top, we compute the similarity between mapped text embeddings and the DINOv2 patches to produce the initial similarity maps. In the bottom part, we produce a background cleaning map for each class derived from the different DINOv2 attention heads. We obtain the final enhanced similarity map of each category through a convex combination of the similarity and background cleaning maps. The output segmentation then results from the final refinement and thresholding steps.

certain about the predicted category and classify these locations as background. However, previous works [23, 51, 52] have introduced custom approaches to improve the capabilities of the model in recognizing the background.

Following this line, we propose a background cleaning procedure, depicted in Fig. 3, that is based on the capabilities of the DINOv2 backbone in focusing on coherent areas and highlighting the foreground through the self-attention heads. Specifically, given N attention maps $A_i \in \mathbb{R}^{\frac{H}{P} \times \frac{W}{P}}$ and M projected textual embeddings of classes $\psi(t_j)$, we first compute the average visual embeddings v^{A_i} as in Eq. 5. Similarly to the training procedure, we then compute the similarity between each v^{A_i} and $\psi(t_j)$, resulting in a matrix of similarity scores $\mathcal{R} \in \mathbb{R}^{M \times N}$, which is additionally normalized row-wise through a softmax operation. These scores represent how much each self-attention head is related to each textual category. Formally, $\mathcal{R}$ is defined as

$$\mathcal{R} = [\mathcal{R}_1, ..., \mathcal{R}_j, ..., \mathcal{R}_M]^\intercal, \text{ with} \quad (7)$$
$$\mathcal{R}_j = \text{softmax}\left(\text{sim}(v^{A_1}, \psi(t_j)), ..., \text{sim}(v^{A_N}, \psi(t_j))\right).$$

Then, for each category T_j we compute its average attention map $\mathcal{F}_j \in \mathbb{R}^{\frac{H}{P} \times \frac{W}{P}}$ as

$$\mathcal{F}_j(A_i, \mathcal{R}_{ij}) = \sum_{i=1}^{N} \mathcal{R}_{ij} A_i, \quad (8)$$

and normalize $\mathcal{F}$ by applying a softmax normalization over both spatial axes, and linearly re-projecting its values in the range $[\min_{j,h,w} \mathcal{S}(I, T_j)_{[h,w]}, \max_{j,h,w} \mathcal{S}(I, T_j)_{[h,w]}]$, where $\mathcal{S}(\cdot)$ is the similarity map defined in Eq. 1. We exploit the resulting normalized average attention per category to shape the similarity map by activating the foreground region and deactivating the background. The resulting shaped similarity map $\overline{\mathcal{S}} \in \mathbb{R}^{H \times W}$ is defined as

$$\overline{\mathcal{S}}(I, T_j)_{[h,w]} = \lambda \mathcal{S}(I, T_j)_{[h,w]} + (1 - \lambda)\mathcal{F}_{j,[h,w]}, \quad (9)$$

where λ is a hyperparameter representing the relevance of the background shaping in computing the segmentation masks. The background mask is then identified as the collection of pixels for which the shaped similarity map is lower than a threshold across all semantic categories.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate our approach on eight widely-used semantic segmentation benchmarks, which we categorize based on the inclusion of a background class. Specifically, we conduct experiments on the validation sets of Pascal VOC 2012 [15], Pascal Context [34], COCO Stuff [7], Cityscapes [11], and ADE20K [61, 62], that contain 20, 59, 171, 150, and 19 semantic categories respectively and do not include the "background" class. We report additional experiments on the COCO Objects dataset [7], which consists of 80 different foreground object classes, and on modified versions of Pascal VOC 2012 and Pascal Context in which the "background" category is, instead, included (*i.e.*, with 21 and 60 semantic categories respectively).

Implementation Details. For the main experiments, we employ DINOv2 ViT-B/14 as the base model and DINOv2 ViT-L/14 as the large model, both with the CLIP ViT-B/16 text encoder. We use the DINOv2 variant with registers [12] since our method benefits from the removal of artifacts in self-attention maps. We train the model with the Adam optimizer, a batch size of 128, and a learning rate of 1×10^{-4} for a total of 100 epochs on the COCO Captions 2014 training split [31], composed of around 80k images.

Following previous works [9, 18, 23], we optionally employ a mask refinement stage to counteract any inaccuracies in the final masks. In particular, we adopt Pixel-Adaptive Mask Refinement (PAMR) [1], an iterative post-refinement method aimed at enhancing the fidelity of the similarities to the visual characteristics of the image. In our experiments, we use λ equal to $5/6$ for background cleaning, a threshold of 0.55 on the similarity score to determine which pixels belong to the "background" category, and, when using mask refinement, employ PAMR with 10 iterations.

Evaluation Protocol. We follow the standard evaluation protocol for unsupervised OVS [9], where prior access to

Model	Visual Encoder	Frozen	ViT-Base (mIoU)									ViT-Large (mIoU)								
			V20	C59	Stuff	City	ADE	V21	C60	Object	**Avg**	V20	C59	Stuff	City	ADE	V21	C60	Object	**Avg**
without Mask Refinement																				
GroupViT [54]	Custom ViT	✗	79.7	23.4	15.3	11.1	9.2	50.4	18.7	27.5	29.4	-	-	-	-	-	-	-	-	-
ReCo [41]	CLIP	✗	57.7	22.3	14.8	21.1	11.2	25.1	19.9	15.7	23.5	-	-	-	-	-	-	-	-	-
TCL [9]	CLIP	✗	77.5	30.3	19.6	23.1	14.9	51.2	24.3	30.4	33.9	-	-	-	-	-	-	-	-	-
SILC [36]	Custom ViT	✗	77.5	31.6	20.8	26.9	19.3	-	-	-	-	-	-	-	-	-	-	-	-	-
MaskCLIP [63]	CLIP	✓	74.9	26.4	16.4	12.6	9.8	38.8	23.6	20.6	27.9	29.4	12.4	8.8	11.5	7.2	23.3	11.7	7.2	13.9
CLIP-DIY [51]	CLIP+DINO	✓	79.7	19.8	13.3	11.6	9.9	59.9	19.7	31.0	30.6	-	-	-	-	-	-	-	-	-
SCLIP [47]	CLIP	✓	80.4	34.2	22.4	32.2	16.1	59.1	30.4	30.5	38.2	70.6	25.2	17.6	21.3	10.9	44.0	22.3	26.9	29.9
CLIP-DINOiser [52]	CLIP	✓	80.9	35.9	24.6	31.1	20.0	**62.1**	32.4	34.8	40.2	-	-	-	-	-	-	-	-	-
ClearCLIP [25]	CLIP	✓	80.9	35.9	23.9	30.0	16.7	51.8	32.6	33.0	38.1	80.0	29.6	19.9	27.9	15.0	-	-	-	-
NACLIP [18]	CLIP	✓	79.7	35.2	23.3	35.5	17.4	58.9	32.2	33.2	39.4	78.7	32.1	21.4	31.4	17.3	52.2	28.7	29.9	36.5
`dino.txt` [21]	DINOv2(reg)	✓	-	-	-	-	-	-	-	-	-	62.1	30.9	20.9	32.1	20.6	-	-	-	-
FreeDA [3]	DINOv2	✓	77.1	37.1	24.9	34.0	19.5	51.7	32.6	24.4	37.7	71.8	35.4	24.2	32.3	19.4	44.9	31.1	24.6	35.5
FreeDA [3]	CLIP+DINOv2	✓	84.3	39.7	25.7	34.1	20.8	51.8	35.3	36.3	41.0	85.7	**39.7**	26.3	33.6	21.4	44.1	34.8	33.9	39.9
ProxyCLIP [26]	CLIP+DINOv2(reg)	✓	83.0	37.2	25.4	33.9	19.7	58.6	33.8	37.4	41.1	85.2	36.2	24.6	35.2	21.6	56.6	33.0	36.7	41.1
ProxyCLIP [26]	CLIP+DINO	✓	80.3	39.1	26.5	**38.1**	20.2	61.3	**35.3**	37.5	42.3	83.2	37.7	25.6	**40.1**	**22.6**	**60.6**	**34.5**	**39.2**	**42.9**
Talk2DINO (Ours)	DINOv2(reg)	✓	**87.1**	**39.8**	**28.1**	36.6	**21.1**	61.5	35.1	**41.0**	**43.8**	**87.1**	39.1	**27.0**	35.8	21.1	60.1	34.2	37.6	42.8
with Mask Refinement																				
GroupViT [54]	Custom ViT	✗	81.5	23.8	15.4	11.6	9.4	51.1	19.0	27.9	30.0	-	-	-	-	-	-	-	-	-
ReCo [41]	CLIP	✗	62.4	24.7	16.3	22.8	12.4	27.2	21.9	17.3	25.6	-	-	-	-	-	-	-	-	-
TCL [9]	CLIP	✗	83.2	33.9	22.4	24.0	17.1	55.0	30.4	31.6	37.2	-	-	-	-	-	-	-	-	-
MaskCLIP [63]	CLIP	✓	72.1	25.3	15.1	11.2	9.0	37.2	22.6	18.9	26.4	-	-	-	-	-	-	-	-	-
SCLIP [47]	CLIP	✓	83.5	36.1	23.9	34.1	17.8	61.7	31.5	32.1	40.1	76.3	27.4	18.7	23.9	11.8	47.8	23.8	26.9	32.1
CLIP-DINOiser [52]	CLIP	✓	81.5	37.1	25.3	31.5	20.6	64.6	33.5	36.1	41.3	-	-	-	-	-	-	-	-	-
LaVG [22]	CLIP+DINO	✓	82.5	34.7	23.2	26.2	15.8	62.1	31.6	34.2	38.3	-	-	-	-	-	-	-	-	-
NACLIP [18]	CLIP	✓	83.0	38.4	25.7	38.3	19.1	64.1	35.0	36.2	42.5	84.5	36.4	24.6	37.1	19.6	57.9	36.4	34.6	41.4
FreeDA [3]	DINOv2	✓	79.5	40.2	27.1	34.4	20.9	52.0	35.2	25.8	39.4	75.2	39.0	27.0	33.1	21.3	45.3	34.3	26.7	37.7
FreeDA [3]	CLIP+DINOv2	✓	85.2	42.1	27.0	33.8	21.8	51.8	37.4	38.6	42.2	87.1	42.4	28.1	33.8	22.6	55.4	37.1	36.1	42.8
ProxyCLIP [26]	CLIP+DINOv2(reg)	✓	83.1	38.9	26.6	35.4	20.3	62.0	35.2	38.7	42.5	85.8	37.6	25.6	37.5	22.5	59.4	34.6	39.0	42.8
ProxyCLIP [26]	CLIP+DINO	✓	80.3	39.4	26.9	**38.6**	20.2	60.8	35.3	37.2	42.3	83.2	38.0	26.2	**41.0**	22.6	60.7	34.7	39.4	43.2
Talk2DINO (Ours)	DINOv2(reg)	✓	**88.5**	**42.4**	**30.2**	38.1	**22.5**	**65.8**	**37.7**	**45.1**	**46.3**	**89.8**	**42.7**	**29.6**	38.4	**22.9**	**66.1**	**37.3**	**42.3**	**46.1**

Table 1. Comparison with unsupervised OVS models on Pascal VOC [15], Pascal Context [34], COCO Stuff [7], COCO Object [7], Cityscapes [11], and ADE20K [61, 62]. For each method, we specify the visual backbone used, along with whether it is frozen or fine-tuned. We report both the variants with and without background for Pascal VOC (V21 and V20) and Pascal Context (C60 and C59). Best results with and without mask refinement are highlighted in bold, overall best results are underlined.

the target data before evaluation is not allowed, and use the default class names provided by the `MMSegmentation` toolbox. The images are resized to have a shorter side of 448, using a sliding window approach with a stride of 224 pixels. All models are evaluated using mean Intersection-over-Union (mIoU) on all the classes of each dataset.

4.2. Comparison with the State of the Art

We compare Talk2DINO with previous state-of-the-art approaches for unsupervised OVS on the five benchmarks that do not include the "background" category and the three benchmarks with the "background" category. We consider as competitors: (i) prototype-based approaches, such as ReCo [41] and FreeDA [3], which aim to create visual prototypes associated with the textual categories, (ii) CLIP adaptations, as MaskCLIP [63], CLIP-DIY [51], SCLIP [47], ClearCLIP [25], and NACLIP [18], which propose architectural modifications to enhance its localization properties, (iii) methods trained on sets of image-caption pairs with objectives designed to force the segmentation capabilities to emerge, like GroupViT [54], TCL [9], SILC [36], and `dino.txt` [21], and (iv) methods that aim to combine the properties of CLIP and DINO, as CLIP-DINOiser [52], LaVG [22], and ProxyCLIP [26].

Table 1 reports the results on the five benchmarks without background (*i.e.*, Pascal VOC-20, Pascal Context-59, COCO Stuff, Cityscapes, and ADE) and the three benchmarks with background (*i.e.*, Pascal VOC-21, Pascal Context-60, and COCO Object). Specifically, we report the performance of both the base and large configurations of both Talk2DINO and the competitors, according to their definitions in the original papers. Moreover, we divide the table into two sections depending on whether a mask refinement technique is employed. Specifically, LaVG exploits a custom region proposer combined with DenseCRF [24], while all the other methods refine their masks with PAMR. Regarding datasets with background, Pascal VOC and COCO Objects present only foreground categories, also referred to as "things" in the literature, while Pascal Context presents both categories from foreground and background, also mentioned as "stuff". Hence, following CLIP-DINOiser [52], we report the performance with

Visual Backbone		mIoU V20	C59	Stuff	City	ADE
DINO	ViT-S	27.7	13.2	7.4	14.8	5.2
DINOv2 (without registers)	ViT-S	83.7	**38.3**	**25.8**	**32.9**	**19.8**
DINOv2 (with registers)	ViT-S	**86.9**	35.3	24.5	27.2	16.9
MAE	ViT-B	9.5	4.3	2.0	4.0	1.1
CLIP	ViT-B	55.4	14.9	12.2	6.2	3.8
DINO	ViT-B	27.3	11.2	7.9	13.6	4.5
DINOv2 (without registers)	ViT-B	74.2	31.9	23.0	27.9	16.5
DINOv2 (with registers)	ViT-B	**87.1**	**39.8**	**28.1**	**36.6**	**21.1**
MAE	ViT-L	6.7	2.7	1.4	4.6	0.9
CLIP	ViT-L	16.6	5.0	7.7	0.9	2.0
DINOv2 (without registers)	ViT-L	56.0	20.1	14.9	18.1	8.2
DINOv2 (with registers)	ViT-L	**87.1**	**39.1**	**27.0**	**35.8**	**21.1**

Table 2. Ablation study results using different visual backbones and with different sizes of the ViT architecture.

	mIoU V20	C59	Stuff	City	ADE
Effect of Projection					
Linear Projection (text only)	85.1	37.9	26.7	35.6	20.1
Our mapping (both vision and text)	59.2	27.3	18.9	23.5	13.5
Our mapping (vision only)	84.6	35.2	26.2	20.4	15.5
Our mapping (text only)	**87.1**	**39.8**	**28.1**	**36.6**	**21.1**
Effect of Self-Attention Selection and Aggregation					
`CLS` only (without self-attention)	84.5	30.6	23.0	22.6	17.2
Standard average	**89.6**	36.9	25.6	33.5	19.7
`CLS`-weighted average	87.6	35.2	23.1	29.3	17.5
`CLS` similarity-weighted sampling	88.2	32.9	22.6	27.0	17.9
Max `CLS` similarity	87.1	**39.8**	**28.1**	**36.6**	**21.1**

Table 3. Ablation study evaluating the impact of the core components of the proposed architecture on the final performance. We report the results using the base model of DINOv2.

the background cleaning procedure described in Sec. 3.3 only on Pascal VOC and COCO Objects.

As it can be observed, our approach achieves the best average mIoU on all the configurations and presents a consistent improvement compared to the considered competitors, with and without the mask refinement, across all datasets except Cityscapes. The most straightforward comparison is the one with FreeDA without global similarity (*i.e.*, with DINOv2 only as visual backbone). It builds a bridge between DINOv2 and the CLIP text encoder by retrieving from a collection of visual-textual embedding pairs and by building visual prototypes for each textual category. The significant improvement achieved by Talk2DINO demonstrates that training a direct projection from the CLIP text encoder to DINOv2 leads to a more accurate bridge between the two embedding spaces without the overhead in computation and memory provided by the retrieval procedure.

4.3. Ablation Studies and Analyses

Choosing Different Visual Backbones. In Table 2 we show the performance of our approach when varying the visual backbone and the size of the employed ViT architecture. We observe that backbones that differ from DINOv2 present unsatisfactory results and can not be aligned to the CLIP textual encoder with a learnable mapping. In particular, while DINO achieves the second best performance on average, Talk2DINO heavily benefits from the strong semantic representation of the dense features of DINOv2 and on the capabilities of its self-attention heads in highlighting coherent regions of the image – properties which are not reflected in other visual backbones. We refer to the supplementary materials for more details on self-attention heads.

Moreover, our results emphasize the critical role of registers [12] in DINOv2, as demonstrated by the comparison between its variants with and without registers. Registers are a recently proposed mechanism to mitigate the presence of artifacts in the feature maps of ViT-based backbones. Artifacts are tokens that exhibit a significantly higher norm with respect to the other tokens and retain less information about their original position in the image. The alignment process in our method relies on high-quality attention maps, and the presence of artifacts poses a challenge by limiting the selection of the most relevant self-attention heads. Interestingly, since register-related artifacts are more pronounced in larger backbones, the ViT-S variant without registers maintains competitive performance compared to its register-enabled counterpart. Finally, we observe that our approach maintains robust and consistent performance across different ViT sizes, achieving strong results even with the compact ViT-S backbone. This suggests that our method is effective across a range of model sizes, making it adaptable to varying computational constraints. Additional insights and detailed evaluations of the different visual backbones can be found in the supplementary.

Impact of the Proposed Components. Table 3 reports the results of Talk2DINO evaluating the impact of its core components on the overall performance. Specifically, in the first section of the table, we analyze the effect of the adopted projection ψ. Replacing it with a linear projection leads to a slight performance drop. The good performance obtained by a linear transform evidences how the DINOv2 and CLIP spaces are intrinsically compatible, as the former can be obtained through an affine transform of the latter without losing too much information. Interestingly, applying the proposed projection on top of DINOv2 or using two projections on both spaces significantly lowers performance, confirming the appropriateness of the proposed approach.

In the second section of the table, we instead study the effect of the selection and aggregation strategy of self-attention heads $A_i, i = 1, ..., N$ during training. In particular, we test aligning (i) directly the visual `CLS` token to the textual `CLS` token, (ii) the visual embedding from the standard average self-attention, (iii) the weighted mean of the head embeddings v_{A_i}, where the weights are given by their

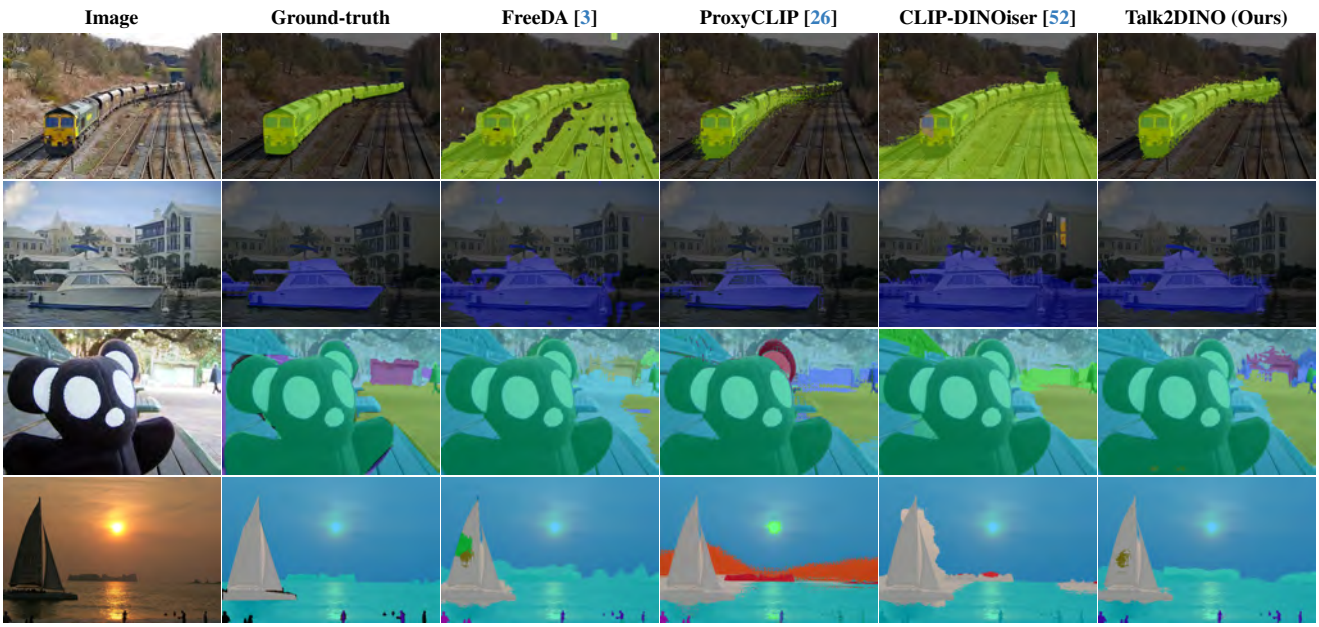

Figure 4. Qualitative results of Talk2DINO in comparison with FreeDA [3], ProxyCLIP [26], and CLIP-DINOiser [52].

	Mask Refinement	mIoU	
		V21	Object
without background cleaning	✗	59.9	37.1
with background cleaning	✗	**61.5**	**41.0**
without background cleaning	✓	63.9	40.3
with background cleaning	✓	**65.8**	**45.1**

Table 4. Ablation study on the impact of the background cleaning procedure. We report the results using DINOv2 ViT-B.

softmaxed similarity with the textual `CLS` token, (iv) a strategy in which we sample a single head embedding, where the sampling probability is given by the softmaxed similarity with the `CLS` token, and (v) our adopted solution in which we select the head embedding which is the most similar to the textual `CLS` token. The results show that only on the Pascal VOC dataset – composed mostly by large subjects in foreground – the embedding from the standard average self-attention presents improved performance. On all other benchmarks, our approach proves to be the most effective, further validating the robustness of our selection method.

Effect of Background Cleaning. Table 4 shows how the performance is affected by the background cleaning mechanism and by the usage of PAMR for mask refinement. It can be observed that the background cleaning procedure has a significantly positive impact on Pascal VOC and COCO Object, leading, respectively, to a +1.6 and +3.9 increase in mIoU score. Further, it can be noticed that the effectiveness of the proposed background cleaning procedure is confirmed also when applying the mask refinement. Qualitative results showing the effect of the background cleaning procedure are available in the supplementary material.

Qualitative Results. Fig. 4 depicts qualitative segmentation results, in which we highlight the segmentation capabilities of Talk2DINO along with other state-of-the-art models (*i.e.*, FreeDA [3], ProxyCLIP [26], and CLIP-DINOiser [52]). We show two images from Pascal VOC, in which it can also be appreciated how the background cleaning procedure leads to high-quality masks and localization, and two images from COCO Stuff and Pascal Context where Talk2DINO effectively segments "things" in the scene, such as the `boat` and `teddy bear`, and "stuff" categories, such as `sky` and `road`.

5. Conclusion

In this paper, we introduced Talk2DINO, a novel approach for OVS that bridges the spatially detailed embeddings of the DINOv2 self-supervised vision backbone with the highly semantic text embeddings of CLIP. Our method achieves fine-grained alignment between textual concepts and visual patches without the need for extensive fine-tuning of the backbone networks, only leveraging self-attention maps from DINOv2 and a lightweight language-to-vision mapping layer. Talk2DINO achieves state-of-the-art results in standard OVS benchmarks, outperforming previous prototype-based, CLIP adaptation, and contrastive learning approaches. Our study highlights the potential of integrating vision-only and multimodal models, suggesting broader applications in fine-grained and cross-modal tasks.

Acknowledgments

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support. This work has been conducted under a research grant co-funded by Leonardo S.p.A., and supported by the PNRRM4C2 project "FAIR - Future Artificial Intelligence Research", by the PRIN 2022-PNRR project "MUCES" (CUP E53D23016290001 and B53D23026090001), and by the PNRR project "IT-SERR - Italian Strengthening of Esfri RI Resilience" (CUP B53C22001770006), all funded by the European Union - NextGenerationEU, and by the SUN XR project funded by the Horizon Europe Research & Innovation Programme (GA n. 101092612).

References

[1] Nikita Araslanov and Stefan Roth. Single-Stage Semantic Segmentation from Image Labels. In *CVPR*, 2020. 5

[2] Luca Barsellotti, Roberto Amoroso, Lorenzo Baraldi, and Rita Cucchiara. FOSSIL: Free Open-Vocabulary Semantic Segmentation through Synthetic References Retrieval. In *WACV*, 2024. 3

[3] Luca Barsellotti, Roberto Amoroso, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation. In *CVPR*, 2024. 1, 3, 6, 8, 2, 9

[4] Ioana Bica, Anastasija Ilic, Matthias Bauer, Goker Erdogan, Matko Bošnjak, Christos Kaplanis, Alexey A Gritsenko, Matthias Minderer, Charles Blundell, Razvan Pascanu, et al. Improving fine-grained understanding in image-text pre-training. In *ICML*, 2024. 2

[5] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding Everything: Emerging Localization Properties in Vision-Language Transformers. In *CVPR*, 2024. 2

[6] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. In *NeurIPS*, 2019. 1

[7] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and Stuff Classes in Context. In *CVPR*, 2018. 4, 5, 6, 3

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *ICCV*, 2021. 2

[9] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning To Generate Text-Grounded Mask for Open-World Semantic Segmentation From Only Image-Text Pairs. In *CVPR*, 2023. 1, 2, 5, 6, 3

[10] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. CAT-Seg: Cost Aggregation for Open-Vocabulary Semantic Segmentation. In *CVPR*, 2024. 2

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *CVPR*, 2016. 5, 6, 3

[12] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision Transformers Need Registers. In *ICLR*, 2024. 2, 5, 7

[13] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *CVPR*, 2022. 2

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021. 2

[15] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, 2012. 4, 5, 6, 3

[16] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA: Exploring the Limits of Masked Visual Representation Learning at Scale. In *CVPR*, 2023. 2

[17] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, 2022. 1, 2

[18] Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay Attention to Your Neighbours: Training-Free Open-Vocabulary Semantic Segmentation. In *WACV*, 2025. 1, 2, 5, 6, 3

[19] Cong Han, Yujie Zhong, Dengjie Li, Kai Han, and Lin Ma. Open-Vocabulary Semantic Segmentation with Decoupled One-Pass Network. In *ICCV*, 2023. 2

[20] 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 *ICML*, 2021. 2

[21] Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, et al. Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24905–24916, 2025. 6

[22] Dahyun Kang and Minsu Cho. In Defense of Lazy Visual Grounding for Open-Vocabulary Semantic Segmentation. In *ECCV*, 2024. 3, 6

[23] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion Models for Zero-Shot Open-Vocabulary Segmentation. In *ECCV*, 2024. 3, 5

[24] Philipp Krähenbühl and Vladlen Koltun. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In *NeurIPS*, 2011. 6

[25] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ClearCLIP: Decomposing CLIP Representations for Dense Vision-Language Inference. In *ECCV*, 2024. 2, 6

[26] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ProxyCLIP: Proxy Attention Improves CLIP for Open-Vocabulary Segmentation. In *ECCV*, 2024. 1, 2, 3, 6, 8, 9

[27] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022. 1

[28] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded Language-Image Pre-Training. In *CVPR*, 2022. 1

[29] Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, 2025. 2

[30] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, 2023. 2

[31] 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 *ECCV*, 2014. 5, 4

[32] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. SegCLIP: Patch Aggregation with Learnable Centers for Open-Vocabulary Semantic Segmentation. In *ICML*, 2023. 2

[33] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling Open-Vocabulary Object Detection. *NeurIPS*, 2024. 1

[34] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The Role of Context for Object Detection and Semantic Segmentation in the Wild. In *CVPR*, 2014. 5, 6, 3

[35] Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip HS Torr, and Ser-Nam Lim. Open Vocabulary Semantic Segmentation With Patch Aligned Contrastive Learning. In *CVPR*, 2023. 2

[36] Muhammad Ferjad Naeem, Yongqin Xian, Xiaohua Zhai, Lukas Hoyer, Luc Van Gool, and Federico Tombari. SILC: Improving Vision Language Pretraining with Self-Distillation. In *ECCV*, 2024. 6

[37] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 4

[38] 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 *ICML*, 2021. 1, 2, 3

[39] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual Grouping in Contrastive Vision-Language Models. In *ICCV*, 2023. 2

[40] Pengzhen Ren, Changlin Li, Hang Xu, Yi Zhu, Guangrun Wang, Jianzhuang Liu, Xiaojun Chang, and Xiaodan Liang. ViewCo: Discovering Text-Supervised Segmentation Masks via Multi-View Semantic Consistency. In *ICLR*, 2023. 2

[41] Gyungin Shin, Weidi Xie, and Samuel Albanie. ReCo: Retrieve and Co-segment for Zero-shot Transfer. In *NeurIPS*, 2022. 3, 6, 1

[42] Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing Objects with Self-Supervised Transformers and no Labels. In *BMVC*, 2021. 2, 3, 4

[43] Oriane Siméoni, Chloé Sekkat, Gilles Puy, Antonín Vobeckỳ, Éloi Zablocki, and Patrick Pérez. Unsupervised Object Localization: Observing the Background To Discover Objects. In *CVPR*, 2023. 2, 3, 4

[44] Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. CLIP as RNN: Segment Countless Visual Concepts without Training Endeavor. In *CVPR*, 2024. 2

[45] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. In *CVPR*, 2024. 2

[46] Weijie Tu, Weijian Deng, and Tom Gedeon. A Closer Look at the Robustness of Contrastive Language-Image Pre-Training (CLIP). In *NeurIPS*, 2023. 2

[47] Feng Wang, Jieru Mei, and Alan Yuille. SCLIP: Rethinking Self-Attention for Dense Vision-Language Inference. In *ECCV*, 2024. 1, 2, 6, 3

[48] Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and Learn for Unsupervised Object Detection and Instance Segmentation. In *CVPR*, 2023. 2, 3

[49] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-Supervised Transformers for Unsupervised Object Discovery Using Normalized Cut. In *CVPR*, 2022.

[50] Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L Crowley, and Dominique Vaufreydaz. TokenCut: Segmenting Objects in Images and Videos With Self-Supervised Transformer and Normalized Cut. *IEEE Trans. PAMI*, 45(12):15790–15801, 2023. 2, 3, 4

[51] Monika Wysoczańska, Michaël Ramamonjisoa, Tomasz Trzciński, and Oriane Siméoni. CLIP-DIY: CLIP Dense Inference Yields Open-Vocabulary Semantic Segmentation For-Free. In *WACV*, 2024. 2, 5, 6

[52] Monika Wysoczanska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzcinski, and Patrick Pérez. CLIP-DINOiser: Teaching CLIP a few DINO tricks for open-vocabulary semantic segmentation. In *ECCV*, 2024. 3, 5, 6, 8, 9

[53] Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. SED: A Simple Encoder-Decoder for Open-Vocabulary Semantic Segmentation. In *CVPR*, 2024. 1, 2

[54] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. GroupViT: Semantic Segmentation Emerges From Text Supervision. In *CVPR*, 2022. 1, 2, 6, 3

[55] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Yi Wang, Yu Qiao, and Weidi Xie. Learning Open-Vocabulary Semantic Segmentation Models From Natural Language Supervision. In *CVPR*, 2023. 2

[56] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, 2023. 2

[57] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A Simple Baseline for Open-Vocabulary Semantic Segmentation with Pre-trained Vision-Language Model. In *ECCV*, 2022. 2

[58] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side Adapter Network for Open-Vocabulary Semantic Segmentation. In *CVPR*, 2023. 1, 2

[59] Hang Zhao, Xavier Puig, Bolei Zhou, Sanja Fidler, and Antonio Torralba. Open Vocabulary Scene Parsing. In *ICCV*, 2017. 1

[60] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. RegionCLIP: Region-Based Language-Image Pretraining. In *CVPR*, 2022. 1, 2

[61] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene Parsing Through ADE20K Dataset. In *CVPR*, 2017. 5, 6, 3

[62] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic Understanding of Scenes Through the ADE20K Dataset. *IJCV*, 127(3):302–321, 2019. 5, 6, 3

[63] Chong Zhou, Chen Change Loy, and Bo Dai. Extract Free Dense Labels from CLIP. In *ECCV*, 2022. 1, 2, 6

[64] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting Twenty-Thousand Classes Using Image-Level Supervision. In *ECCV*, 2022. 1

[65] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized Decoding for Pixel, Image, and Language. In *CVPR*, 2023. 1, 2

Talking to DINO: Bridging Self-Supervised Vision Backbones with Language for Open-Vocabulary Segmentation

Supplementary Material

In the following, we present additional material and experimental analyses of the proposed Talk2DINO approach.

A. Additional Experiments and Analyses

Analysis of Model Parameters. Fig. 5 reports a comparison of the relationship between the performance, in terms of average mIoU, and the number of parameters of the models. As it can be observed, Talk2DINO presents a lower number of parameters than the recent competitors FreeDA [3] and ProxyCLIP [26], along with an improved average mIoU. Models with a comparable number of parameters, such as TCL [9], GroupViT [54], and MaskCLIP [63], exhibit a lower performance compared to Talk2DINO. Finally, it shall be noted that models such as FreeDA and ReCo [41] require maintaining external sources of knowledge, which increases memory consumption. Further discussion on the comparison between Talk2DINO, ProxyCLIP, and FreeDA can be found in the following sections (see "*Comparison with ProxyCLIP and FreeDA*").

Role of DINO Registers. The main configuration of Talk2DINO, with both the base and large sizes, leverages the variant of DINOv2 with registers. In Fig. 8 we depict, on the first row, the average self-attentions between the `CLS` and the other tokens for the ViT-S, ViT-B, and ViT-L architectures with and without registers, while in the following rows, we show the various self-attention heads for each backbone. It can be observed that in the ViT-S the artifacts are not present, and the average self-attention between the model with and without the registers is nearly identical. Instead, the ViT-B exhibits artifacts in the top left corner, resulting in an average self-attention that is especially focused on that portion of the image. This side effect is even more noticeable with the ViT-L, for which the artifact is the only visible token in the average self-attention. These observations align with the results reported in Tab. 2, that show a downgrade in performance without the registers that is directly related to the presence of the artifacts in the self-attentions. Indeed, the largest difference in performance is measured in the ViT-L architecture, while in the ViT-S case, the backbone without registers performs better on four benchmarks out of five.

Effect of Training CLIP Last Layer. Table 6 reports a comparison between Talk2DINO when training only the $\psi(t)$ projection as proposed in the main paper and when instead unfreezing the last layer of CLIP [38]. Despite this experiment exhibiting a small performance gap between the two configurations, unfreezing the last layer of CLIP, interestingly, leads to worse results. This outcome highlights that the textual representations provided by CLIP, which have been pre-trained to match their visual counterpart, if trained inside a different pipeline, can be harmed and can lose part of their capabilities in multimodal understanding.

Choosing Different Visual backbones. In Table 2 of the main paper, we report the performance of our approach

Figure 5. **Performance vs. Parameter Count.** The y-axis denotes the obtained mIoU averaged over all the five benchmarks reported in the main paper. Dashed lines connect methods tested with multiple backbone sizes. These are labeled S/B/L for Small/Base/Large ViT models. Talk2DINO offers the best trade-off between performance and number of parameters.

Visual Backbone	ViT-S	ViT-B	ViT-L
MAE	-	0.56	0.56
CLIP	-	0.89	0.89
DINO	0.62	0.73	-
DINOv2 (without registers)	0.95	0.96	0.95
DINOv2 (with registers)	**0.96**	**0.97**	**0.96**

Table 5. Patch linear probing accuracy on VOC for ViT-Small, ViT-Base, and ViT-Large.

applied to different visual backbones. The results demonstrated that our training pipeline is especially suitable for DINOv2, while it leads to unsatisfactory performance on DINO, MAE, and CLIP. We attribute this performance gap to two major factors: (i) the quality of the attention maps and (ii) the semantic richness of the patch representations.

For the first point, we qualitatively analyze the self-attention patterns of the different backbones. Fig. 9 showcases the average self-attentions between the CLS token and the other tokens in the first row, breaking down the contributions from the various self-attention heads in the successive rows. We observe that the self-attention heads of CLIP introduce a noise pattern similar to what we observed for DINOv2 without registers, which limits the effectiveness of our training pipeline. On the other hand, the self-attention maps of DINO and MAE appear cleaner and emphasize homogeneous image regions. However, in these cases, the performance gap with DINOv2 can be attributed to the insufficient semantic richness of the extracted dense features.

To quantitatively assess the patch-level semantics of these backbones, we conduct an experiment in which we classify each patch through linear probing on the images of VOC. We determine the ground-truth labels of the patches via majority voting and evaluate accuracy on the validation set (batch size = 16, learning rate = 5×10^{-3}, for 3 epochs, with 32×32 patches per image, using ViT-B as the backbone). The results, reported in Table 5, align with the overall trends highlighted in the paper: DINOv2 consistently emerges as the best-performing backbone, MAE as the worst, and CLIP and DINO as intermediate. These findings further confirm that the semantic richness of the features extracted by different backbones plays a crucial role in the effectiveness of our approach. Similar conclusions were drawn in the ablation study of FreeDA [3], where a comparable performance drop was observed when using CLIP or DINO instead of DINOv2.

Using CLIP Text Tokens. In Table 6, we report the results of utilizing the dense output of the CLIP text encoder instead of its CLS token for alignment. While our primary experiments align the CLS token with the best attention map embedding to target the patches most relevant to the text, we also explore aligning individual text tokens to the best attention map embeddings. This approach is motivated by the hypothesis that each word in the text might correspond

	mIoU				
	V20	C59	Stuff	City	ADE
Effect of Training CLIP Last Layer					
Trained	77.9	31.5	21.3	34.6	18.7
Frozen	**87.1**	**39.8**	**28.1**	**36.6**	**21.1**
Effect of Text Token Selection					
Average text tokens	84.7	37.9	25.7	33.6	20.0
Text token to best self-attn map	83.9	33.8	24.2	29.5	18.1
Text token to best self-attn map (NS)	80.8	33.9	23.7	27.5	18.6
CLS token only	**87.1**	**39.8**	**28.1**	**36.6**	**21.1**

Table 6. Ablation study on the effect of training the last layer of CLIP and text token selection strategies.

to a distinct region in the image. During inference, since we perform the alignment on individual text tokens rather than the CLS token, we average the text tokens to calculate similarity with the visual patches. However, this method yields inferior results compared to using the CLS token. We then refine this approach by aligning only a subset of text tokens selected using nucleus sampling ($\alpha = 0.6$) to filter out potentially irrelevant words, such as stop words. Despite this effort, performance does not improve.

These observations suggest that the global objective of the training of CLIP, similar to its effect on visual patch embeddings, may not endow text tokens with strong local properties that accurately reflect the specific word each embedding represents. This limitation likely contributes to the noisiness of such alignments. Additionally, we evaluate the use of the average of CLIP text tokens in both training and inference as an alternative to the CLS token. While this approach slightly improves over aligning individual tokens, it still underperforms compared to the CLS token, indicating that it encapsulates the most useful and less noisy information for alignment with DINOv2 patches.

Impact of Image Resolution. According to the evaluation protocol introduced in GroupViT [54] and standardized in TCL [9], the images are resized to have a shorter side of 448, and a sliding window approach with a stride of 224 pixels is employed. However, Wang *et al*. [47] observed that the approaches based on CLIP benefit from employing a shorter side of 336 with 224 × 224 windows and a stride of 112 pixels, leading to an equivalent computational effort but better performance. This phenomenon is attributed to two reasons: (i) each window has the same resolution on which CLIP has been originally trained, and (ii) CLIP presents an impressive global understanding but lacks localization capabilities, hence relying on many smaller windows is more advantageous than more patches. This variation of the evaluation setting is not necessary for Talk2DINO because DINOv2, which is the frozen underlying visual encoder, has been trained with a 518 × 518 resolution and presents an outstanding patch-level understanding, However, Tab. 7 reports the results obtained following the setting used in

Model	Visual Backbone	Resolution	ViT-Base (mIoU)						ViT-Large (mIoU)					
			V20	C59	Stuff	City	ADE	**Avg**	V20	C59	Stuff	City	ADE	**Avg**
without Mask Refinement														
SCLIP [47]	CLIP	336	80.4	34.2	22.4	32.2	16.1	37.1	70.6	25.2	17.6	21.3	10.9	29.1
NACLIP [18]	CLIP	336	79.7	35.2	23.3	35.5	17.4	38.2	78.7	32.1	21.4	31.4	17.3	36.2
ProxyCLIP [26]	CLIP+DINOv2	336	80.5	37.3	25.3	35.8	19.0	39.6	83.5	36.7	25.0	35.8	21.0	40.4
ProxyCLIP [26]	CLIP+DINO	336	78.2	38.8	26.2	**39.7**	19.7	40.5	82.1	**38.2**	**26.2**	**41.2**	**22.2**	**42.0**
Talk2DINO (Ours)	DINOv2	336	**88.3**	**39.1**	**27.4**	38.2	**20.2**	**42.6**	**86.6**	**38.2**	26.0	36.4	19.3	41.3
with Mask Refinement														
SCLIP [47]	CLIP	336	79.3	34.6	22.3	20.3	15.4	34.4	66.6	22.4	14.7	6.9	7.7	23.7
NACLIP [18]	CLIP	336	83.0	38.4	25.7	38.3	19.1	40.9	84.5	36.4	24.6	37.1	19.6	40.4
ProxyCLIP [26]	CLIP+DINOv2	336	80.9	39.3	26.6	37.7	19.9	40.9	83.5	36.7	26.4	38.6	22.1	41.5
ProxyCLIP [26]	CLIP+DINO	336	78.5	39.3	26.7	40.1	20.0	40.9	82.6	38.7	26.7	**42.1**	**22.5**	42.5
Talk2DINO (Ours)	DINOv2	336	**89.4**	**41.5**	**29.4**	**40.3**	**21.2**	**44.4**	**89.5**	**41.7**	**29.8**	38.7	20.8	**44.1**
without Mask Refinement														
SCLIP [47]	CLIP	448	77.8	33.0	21.1	19.8	14.6	33.3	61.2	20.5	13.1	6.7	7.0	21.7
NACLIP [18]	CLIP	448	71.3	34.8	22.9	33.7	17.7	36.1	74.5	32.6	21.6	30.5	17.8	35.4
ProxyCLIP [26]	CLIP+DINOv2	448	83.3	37.8	25.6	28.8	19.1	38.9	85.0	36.6	25.0	33.8	20.6	40.2
ProxyCLIP [26]	CLIP+DINO	448	80.4	39.0	26.2	31.7	19.5	39.4	83.1	37.8	25.9	**37.5**	**21.6**	41.2
Talk2DINO (Ours)	DINOv2	448	**87.1**	**39.8**	**28.1**	**36.6**	**21.1**	**42.5**	**87.1**	**39.1**	**27.0**	35.8	21.1	**42.0**
with Mask Refinement														
SCLIP [47]	CLIP	448	79.3	34.6	22.3	20.3	15.4	34.4	66.6	22.4	14.7	6.9	7.7	23.7
NACLIP [18]	CLIP	448	74.9	37.6	25.2	36.1	18.4	38.4	79.8	36.8	25.0	35.6	18.4	39.1
ProxyCLIP [26]	CLIP+DINOv2	448	83.1	39.3	26.7	29.5	19.7	39.7	85.1	37.8	25.9	35.3	21.4	41.1
ProxyCLIP [26]	CLIP+DINO	448	80.0	39.1	26.5	31.7	19.5	39.4	82.8	37.8	26.2	26.2	21.6	38.9
Talk2DINO (Ours)	DINOv2	448	**88.5**	**42.4**	**30.2**	**38.1**	**22.5**	**44.3**	**89.8**	**42.7**	**29.6**	**38.4**	**22.9**	**44.7**

Table 7. Comparison with unsupervised OVS models on Pascal VOC [15], Pascal Context [34], COCO Stuff [7], Cityscapes [11], and ADE20K [61, 62] following the evaluation setting proposed in SCLIP [47] (*resolution 336*) and TCL [9] (*resolution 448*).

SCLIP, employing a shorter side of 336 for VOC, Context, COCO-Stuff and ADE, of 560 for Cityscapes, with 224 $\times$ 224 windows and stride 112. Results show that Talk2DINO, on average, performs better with a resolution of 448, but the performance slightly varies when changing the setting to 336. This confirms that the semantics of the patch-level features of DINOv2 are robust towards variations of resolution and that our learned bridge is valid for both scenarios. Moreover, for a fair comparison, we also report the results of SCLIP, NACLIP, and ProxyCLIP when adopting the 448 resolution of the standard protocol, in which Talk2DINO largely outperforms the competitors.

Comparison with ProxyCLIP and FreeDA. GroupViT [54] has been the first model to tackle the weakly-supervised OVS. It trains a custom ViT architecture from scratch by hierarchically merging tokens at different layers. Afterward, several works followed this direction, investigating how to let the segmentation capabilities to emerge by training over a large corpora of image-caption pairs. On the contrary, more recent works focused on finding modifications to the architecture of CLIP in order to improve its localization properties. Moreover, some methods consider the usage of further visual encoders with enhanced localization capabilities to help CLIP on dense tasks. Among these methods, ProxyCLIP and FreeDA study how to combine DINO and DINOv2 with CLIP.

FreeDA employs Stable Diffusion to create a huge collection of *localized* images from captions, detecting the area in which each noun of the caption has been generated. This information is used to build a database of textual-visual embedding pairs, in which the textual embedding is obtained with CLIP on each noun and the visual embedding is the average patch-level embedding of DINOv2 from the corresponding area. Then, at inference time, a set of textual embeddings is retrieved for each input category, and the corresponding visual embeddings are averaged to create a prototype for that category in the space of DINOv2. Finally, the CLIP visual encoder runs on the input image to solve ambiguities and remove noise. ProxyCLIP proposes to leverage the semantic coherence of a visual encoder such as DINO or DINOv2 to guide the computation of the patch-level embeddings of CLIP. This guidance is performed inside an attention module, in which the patch-level embeddings of DINO act as queries and keys while those of CLIP act as values.

Talk2DINO, similarly to FreeDA and ProxyCLIP, investigates how to leverage DINOv2 to compensate for CLIP. However, we propose to employ contrastive learning over a large set of image-caption pairs based on maximum similarity between the attention head embeddings and texts, to learn a functional mapping that bridges the CLIP text embeddings into the DINOv2 space. Our approach demonstrates that the two spaces can be directly connected to set

	Image→Text					Text→Image				
	R@1 ↑	R@5 ↑	R@10 ↑	Median ↓	Mean ↓	R@1 ↑	R@5 ↑	R@10 ↑	Median ↓	Mean ↓
ViT-Base										
CLIP	**41.3**	**65.8**	**76.3**	**2**	13.4	22.6	44.1	54.9	8	52.5
Talk2DINO	29.5	56.0	69.0	4	16.4	12.5	34.0	48.4	11	38.4
+ Custom Alignment	28.6	58.8	72.0	4	**12.0**	**28.0**	**55.6**	**68.7**	**4**	**19.3**
ViT-Large										
CLIP	**45.4**	**71.1**	**79.2**	**2**	**11.0**	**26.5**	48.7	59.0	6	44.2
Talk2DINO	26.5	53.7	65.6	5	18.8	12.7	33.7	47.8	11	43.1
+ Custom Alignment	37.9	64.7	75.1	3	13.1	24.4	**50.1**	**63.2**	**5**	**27.8**

Table 8. Retrieval performance on the COCO Captions test set.

	Visual Encoder	Params (M)	FLOPS (G)	Ext. (GiB)
ProxyCLIP	CLIP ViT-B/16 + DINO ViT-B/8	172.0	521.2	-
ProxyCLIP	CLIP ViT-B/16 + DINOv2 ViT-B/14	172.8	180.8	-
FreeDA	CLIP ViT-B/16 + DINOv2 ViT-B/14	172.8	125.1	12.5
Talk2DINO	DINOv2 ViT-B/14	86.6	107.4	-

Table 9. Number of parameters, FLOPS, and the dimension of the external knowledge for ProxyCLIP, FreeDA, and Talk2DINO.

the new state-of-the-art in the unsupervised OVS field. Table 9 shows a quantitative comparison in terms of the number of parameters and FLOPS of the visual encoders and the dimension of the external knowledge (*i.e.*, the database of FreeDA), when assuming an input image with a resolution of 448 $\times$ 448. The results highlight that our method is more practical and less demanding in computation and memory, while presenting improved results against all competitors.

In Tab. 1, we followed the original configurations of the competitors and, hence, ProxyCLIP uses DINOv2 with registers while FreeDA does not. We report a comparison with and without registers in Tab. 10. The registers present the greatest impact on Talk2DINO, because, as described in "*Role of DINO Registers*", the presence of anomaly tokens leads all the self-attention heads to focus only on them, preventing the selection of diverse areas during training and, hence, limiting the efficacy of our proposal. Moreover, in Tab. 10 we report the effect of the background cleaning also on FreeDA and ProxyCLIP. This approach is effective only on Talk2DINO due to the learned alignment between text and average embeddings of the self-attention heads, while it leads to lower results when applied to the other methods.

ViT-B vs ViT-L. Tab. 1 of the main paper shows that, without mask refinement, the results achieved by Talk2DINO with DINOv2 ViT-B as vision encoder are slightly better than the ones achieved with ViT-L, while the opposite should be expected. However, when we apply the PAMR for mask refinement, the results of ViT-L significantly improve, surpassing the ViT-B on five benchmarks out of eight. A similar phenomenon can be observed in other competitors, such as MaskCLIP, SCLIP, ClearCLIP, and NACLIP, while in FreeDA and ProxyCLIP we cannot establish an encoder size that prevails on the other. Even from the experiment in Tab. 5 on patch-level linear probing, we can observe that ViT-B performs slightly better than ViT-L. These results suggest that DINOv2 ViT-L has a comparable semantic understanding with respect to ViT-B, but presents inferior localization properties, which are compensated through PAMR. We hypothesize that training the model with a form of weak- or self-supervision by exploiting the innate capabilities of pre-trained backbones lacks a direct relation between performance and model size. Indeed, the impressive semantic and localized understanding of DINOv2 is a consequence of its training procedure but not the direct objective. From Figure 8, it is noteworthy that the activations of ViT-S, ViT-B, and ViT-L have very different behaviors, impacting the results of Talk2DINO.

B. Image-Text Matching Results

While Talk2DINO is primarily designed for OVS, we also assess its performance on image-text retrieval to evaluate its capabilities in global image understanding. For this task, we adopt the same text encoding approach used in segmentation, projecting the CLIP text embedding. The global image representation is derived by averaging the embeddings computed from each DINOv2 attention map. Specifically, for each attention map A_i, we calculate a visual embedding $v^{A_i} \in \mathbb{R}^{D_v}$ as the weighted average of the dense feature map v. The final global image representation is then obtained by taking the mean of all v^{A_i} embeddings.

In Table 8, we assess the retrieval performance on the COCO Captions test set [31] using both ViT-B and ViT-L configurations. While Talk2DINO generally performs slightly below CLIP across most metrics, it demonstrates a notable advantage in the mean rank for the text-to-image retrieval task. This result underscores the ability of Talk2DINO to better address extreme failures compared to CLIP, indicating improved robustness in handling

Model	Visual Encoder	V20	C59	Stuff	City	ADE	V21	C60	Object	Avg
DINOv2 ViT-B/14 with registers (without Mask Refinement)										
FreeDA	DINOv2	83.4	39.5	25.9	35.2	20.7	50.1 ▷ 43.6	34.3	23.8 ▷ 24.7	39.1 ▷ 38.4
FreeDA	CLIP+DINOv2	87.0	**40.6**	25.7	34.2	**21.2**	49.3 ▷ 41.8	**35.7**	34.8 ▷ 34.7	41.1 ▷ 40.1
ProxyCLIP	CLIP+DINOv2	83.0	37.2	25.4	33.9	19.7	58.6 ▷ 60.0	33.8	37.4 ▷ 37.3	41.1 ▷ 41.3
Talk2DINO	DINOv2	**87.1**	39.8	**28.1**	**39.6**	21.1	59.9 ▷ **61.5**	35.1	37.1 ▷ **41.0**	43.5 ▷ **44.2**
DINOv2 ViT-B/14 with registers (with Mask Refinement)										
FreeDA	DINOv2	84.9	42.3	27.7	36.8	22.0	50.2 ▷ 43.7	36.7	24.5 ▷ 25.5	40.6 ▷ 40.0
FreeDA	CLIP+DINOv2	87.4	**42.4**	26.6	34.8	22.1	49.4 ▷ 41.7	37.2	36.6 ▷ 36.7	42.1 ▷ 41.1
ProxyCLIP	CLIP+DINOv2	83.1	38.9	26.6	35.4	20.3	62.0 ▷ 63.4	35.2	38.7 ▷ 38.6	42.5 ▷ 42.7
Talk2DINO	DINOv2	**88.5**	**42.4**	**30.2**	**41.6**	**22.5**	63.9 ▷ **65.8**	**37.7**	40.3 ▷ **45.1**	45.9 ▷ **46.7**

Table 10. Comparison between FreeDA, ProxyCLIP, and Talk2DINO when using DINOv2 with and without registers. For VOC21, Object, and the average, we report the results without background cleaning on the left and with background cleaning on the right.

challenging or outlier queries. In addition to computing text-image similarities using cosine similarity between a global text token and a global image token, we experiment with a similarity function that mirrors the one used during training. Specifically, instead of representing the image with the mean of the v^{A_i} embeddings and calculating similarity as the cosine similarity between this representation and the text encoding, we represent the image using all v^{A_i} embeddings. We compute the similarity as $\max_{i=1,\dots,N} \text{sim}(v^{A_i}, t)$, taking the maximum similarity score across all heads. This alternative similarity function leads to significant performance improvements, allowing Talk2DINO to surpass CLIP on several metrics. This enhancement is likely due to the ability of the model to evaluate captions at a finer granularity. Captions often describe multiple aspects of an image, including both foreground and background elements. By individually examining different regions of the image as detected by distinct attention heads, the model can assign more precise scores, ultimately boosting retrieval accuracy.

C. Activation Map Visualizations

In Fig. 6, we show the distribution of attention heads selected for alignment with the text input during the final epoch of training. The results indicate that certain heads, particularly heads 1 and 3, are more often aligned with the text than others. However, aside from these, the remaining heads are relatively evenly distributed. These findings are noteworthy because they suggest that some heads specialize in capturing features that align more closely with the input caption, while all heads contribute meaningfully during training. Notably, no head shows a negligible activation frequency, highlighting the importance of the entire set of attention heads in the alignment process.

Fig. 7 presents examples from the training set, showcasing images paired with their corresponding captions and the attention maps selected for alignment. Despite describing the same scene, variations in the captions lead the alignment procedure to focus on different regions of the image. For instance, in the first row, the caption mentioning the fans also

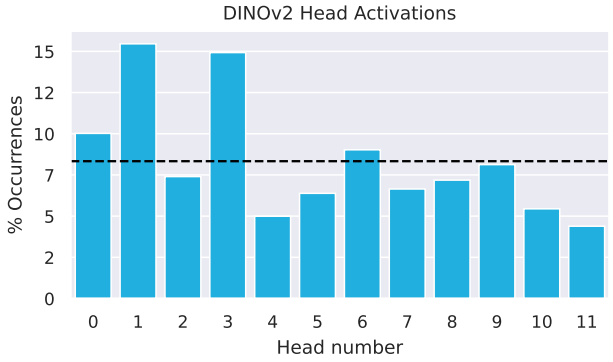

Figure 6. Percentage of times each attention head of ViT-B backbone is selected for alignment to textual embeddings on the final epoch of training. The dashed line denotes uniform distribution.

focuses on the background, while captions that reference only the player, the ball, and the racket do not.

D. Additional Qualitative Results

Effect of Background Cleaning. Fig. 10 shows a set of qualitative results in which we highlight the advantages of using the proposed background cleaning procedure with respect to directly thresholding the similarities with the input categories to detect the background. In particular, the first two rows show four qualitatives on images from COCO Object and the last two rows from VOC. These results demonstrate that background cleaning removes the noise in the background from the image and improves the fitting of the masks on the foreground objects. These findings are reflected in the results reported in Table 4 of the main paper.

In-the-Wild Qualitative Examples. Fig. 11 depicts a few examples of "in-the-wild" segmentation, obtained by providing to Talk2DINO sample images from the web and asking it to segment uncommon categories, such as "`pikachu`", "`millennium falcon`", and "`westminster abbey`", and free-form text, like "`golden retriever puppy`". On the left, we show three examples in which we task the model with also finding the background, while exploiting the background clean-

Figure 7. Sample images from the training set paired with their corresponding captions and the attention maps selected for alignment during the last epoch of training.

ing procedure, and, on the right, three examples in which the model has to assign a provided category to each pixel. The high quality of the resulting masks demonstrates the efficacy of our approach, even on out-of-domain images. From these examples, we can appreciate the capabilities of the model in combining the knowledge from CLIP with the semantic localization of DINOv2 on unconventional concepts, such as fictional character names and proper nouns of historical buildings.

Comparison with State-of-the-Art Methods. Finally, in Fig. 12 we report a set of qualitative results on the five datasets used for the evaluation of the models, in addition to the qualitative depicted in Fig. 4 of the main paper. We compare the segmentation masks of Talk2DINO with the ones of FreeDA [3], ProxyCLIP [26], and CLIP-DINOiser [52], which represent our main competitors. In particular, we report a pair of images from Pascal VOC with background and eight pairs of images from Pascal Context, COCO Stuff, Cityscapes, and ADE20K, without background. As it can be seen, these qualitative results further highlight the impressive segmentation capabilities of Talk2DINO with both background and foreground categories.

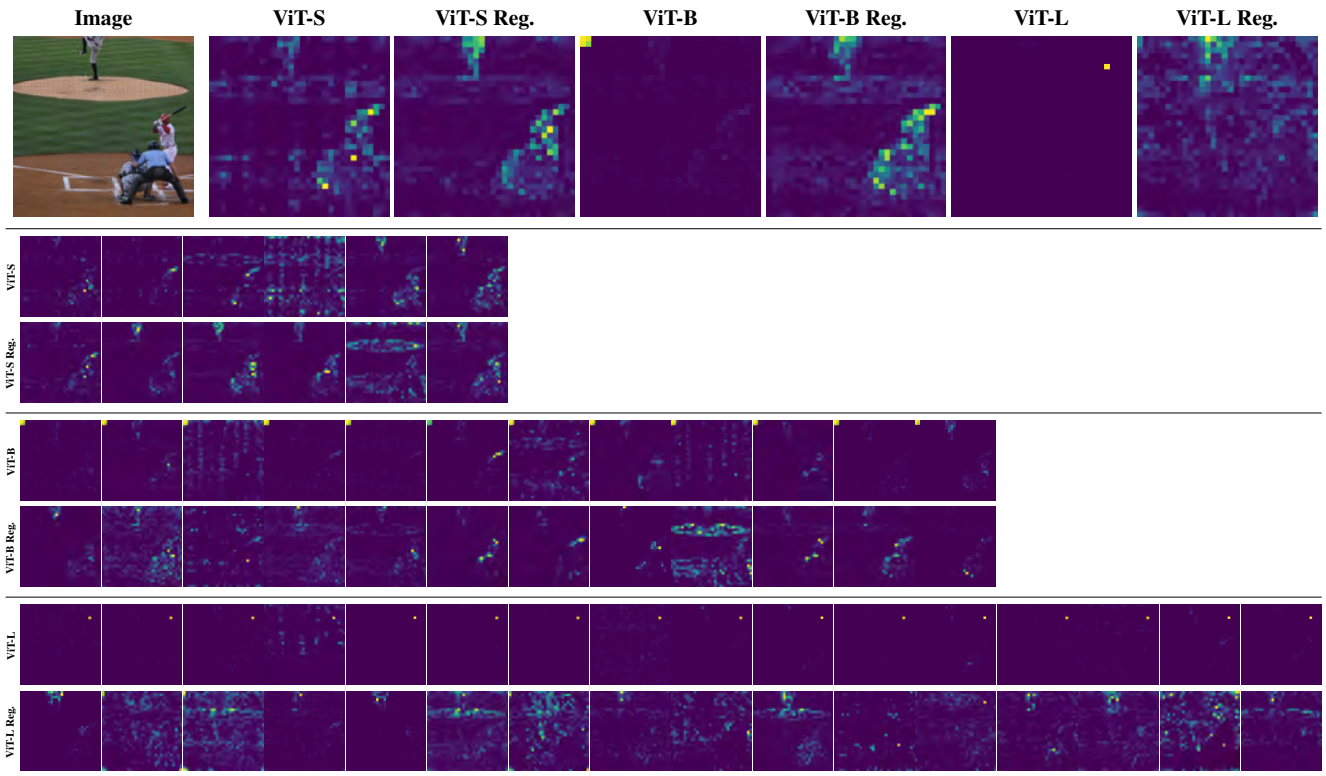

Figure 8. Comparison of DINOv2 with and without registers across different visual backbones (ViT-S, ViT-B, and ViT-L). The results highlight how the ViT-B and ViT-L backbones without registers exhibit artifacts that introduce noise during the alignment process.

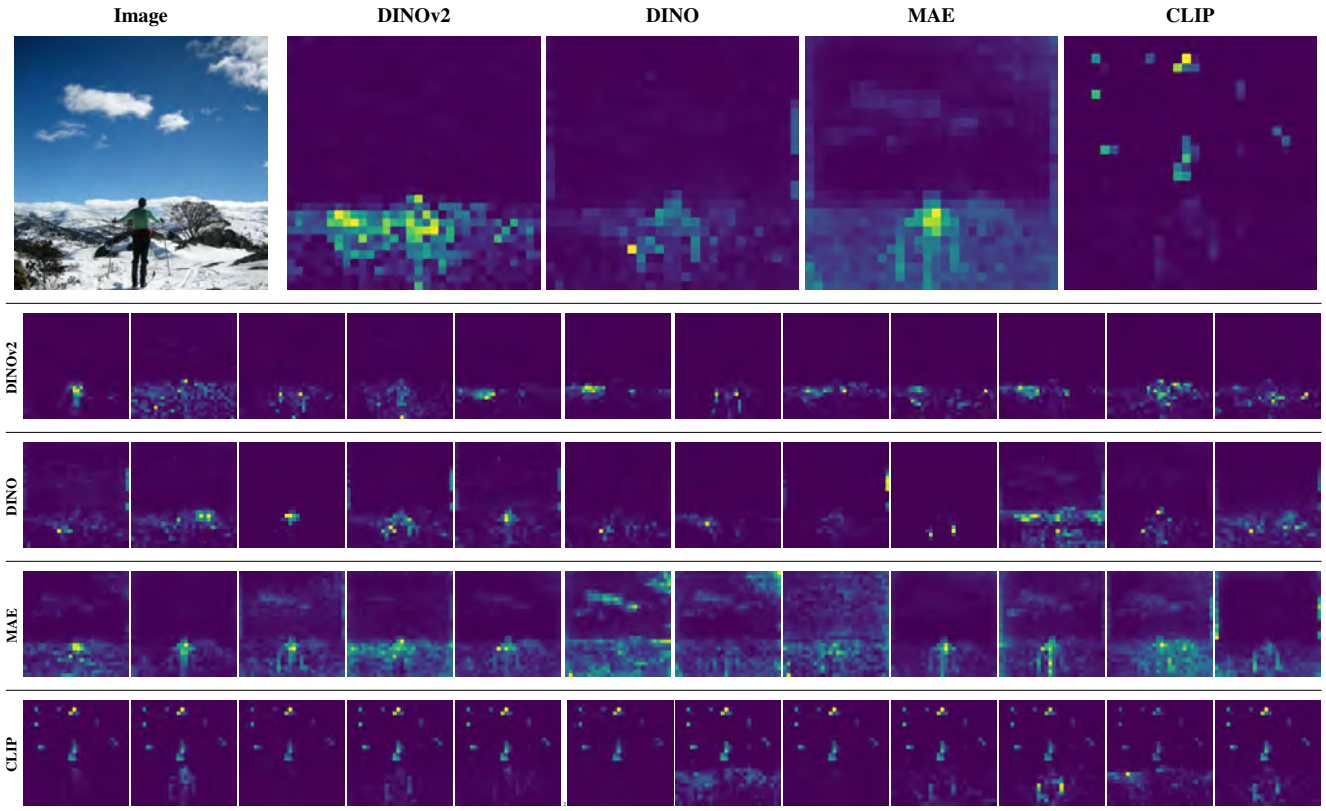

Figure 9. Self-attention activations of different visual backbones (*i.e.*, DINOv2, DINO, MAE, CLIP).

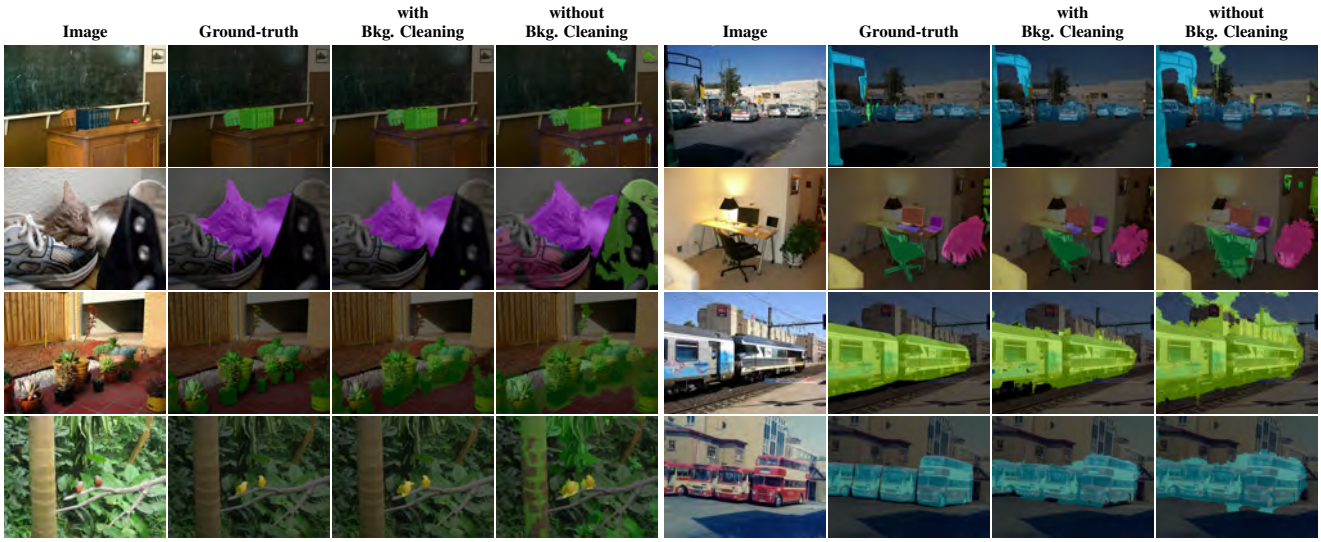

Figure 10. Qualitative results obtained with and without the proposed background cleaning strategy, on COCO Object and Pascal VOC.

Figure 11. ”In-the-wild” segmentation results obtained by prompting Talk2DINO with uncommon textual categories on images retrieved from the web.

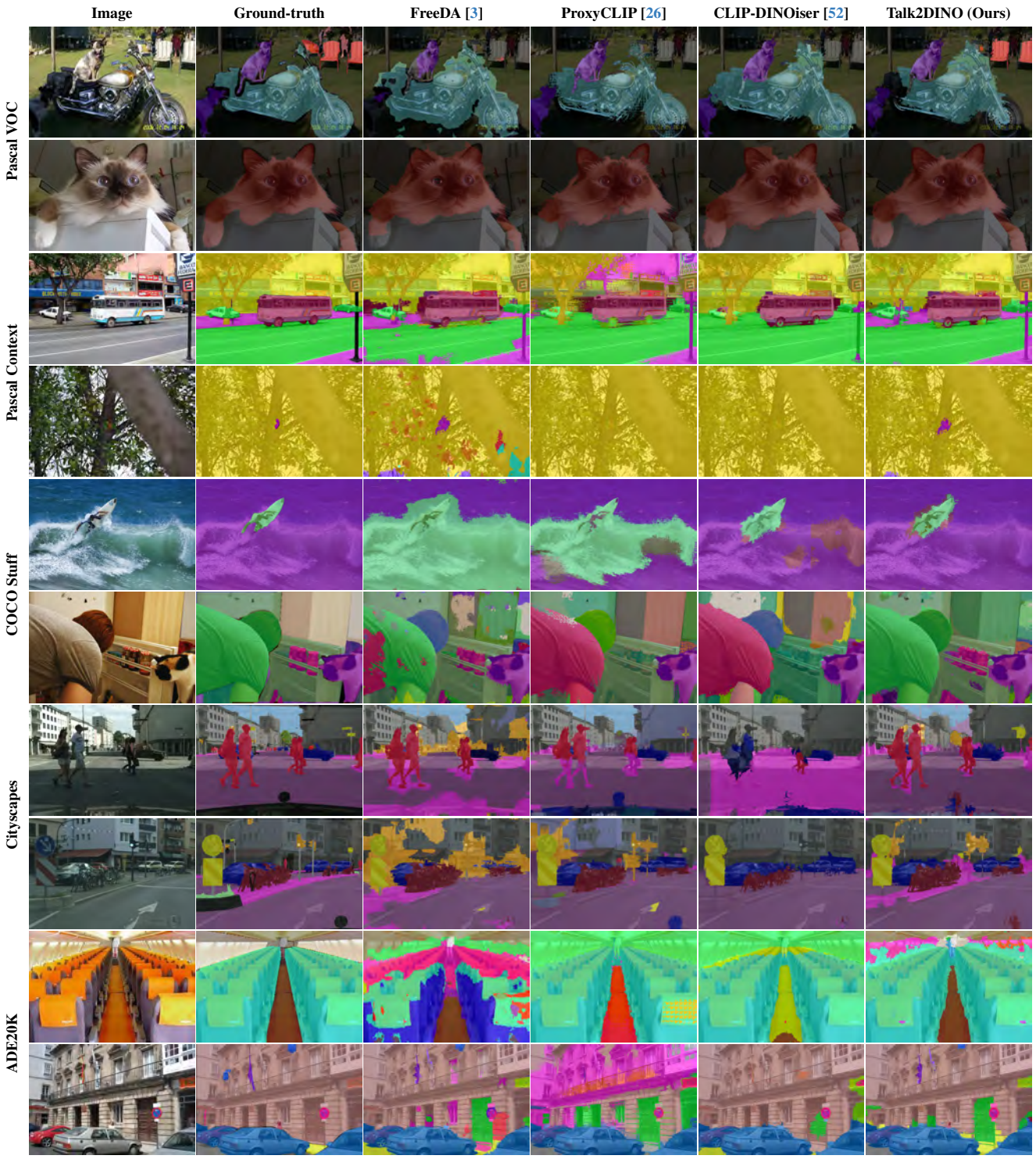

Figure 12. Additional qualitative results of Talk2DINO in comparison with FreeDA [3], ProxyCLIP [26], and CLIP-DINOiser [52].