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

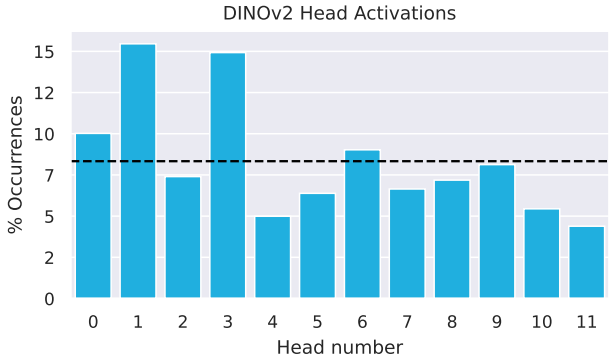

Figure 6. Percentage of times each attention head of ViT-B backbone is selected for alignment to textual embeddings on the final epoch of training. The dashed line denotes uniform distribution.

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.

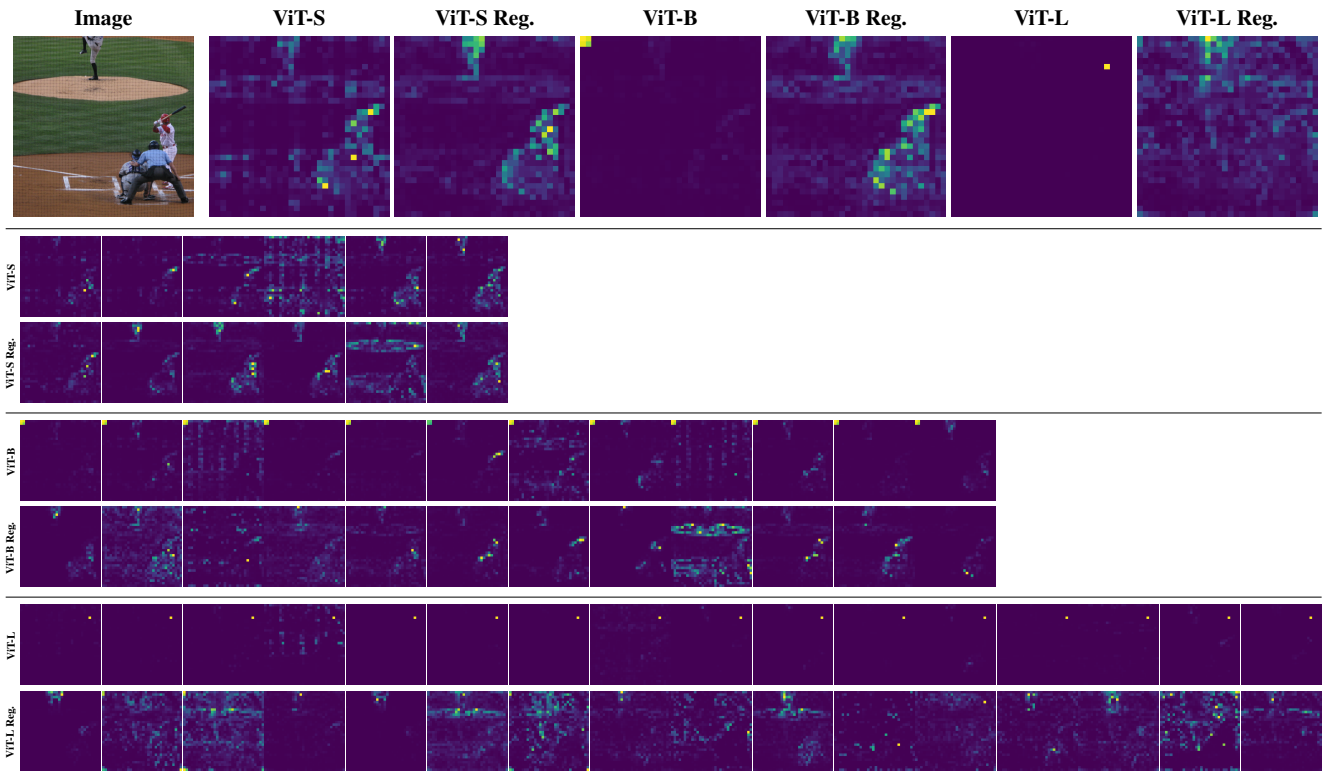

Figure 8. Comparison of DINOv2 with and without registers across different visual backbones (ViT-S, ViT-B, and ViT-L). The results highlight how the ViT-B and ViT-L backbones without registers exhibit artifacts that introduce noise during the alignment process.

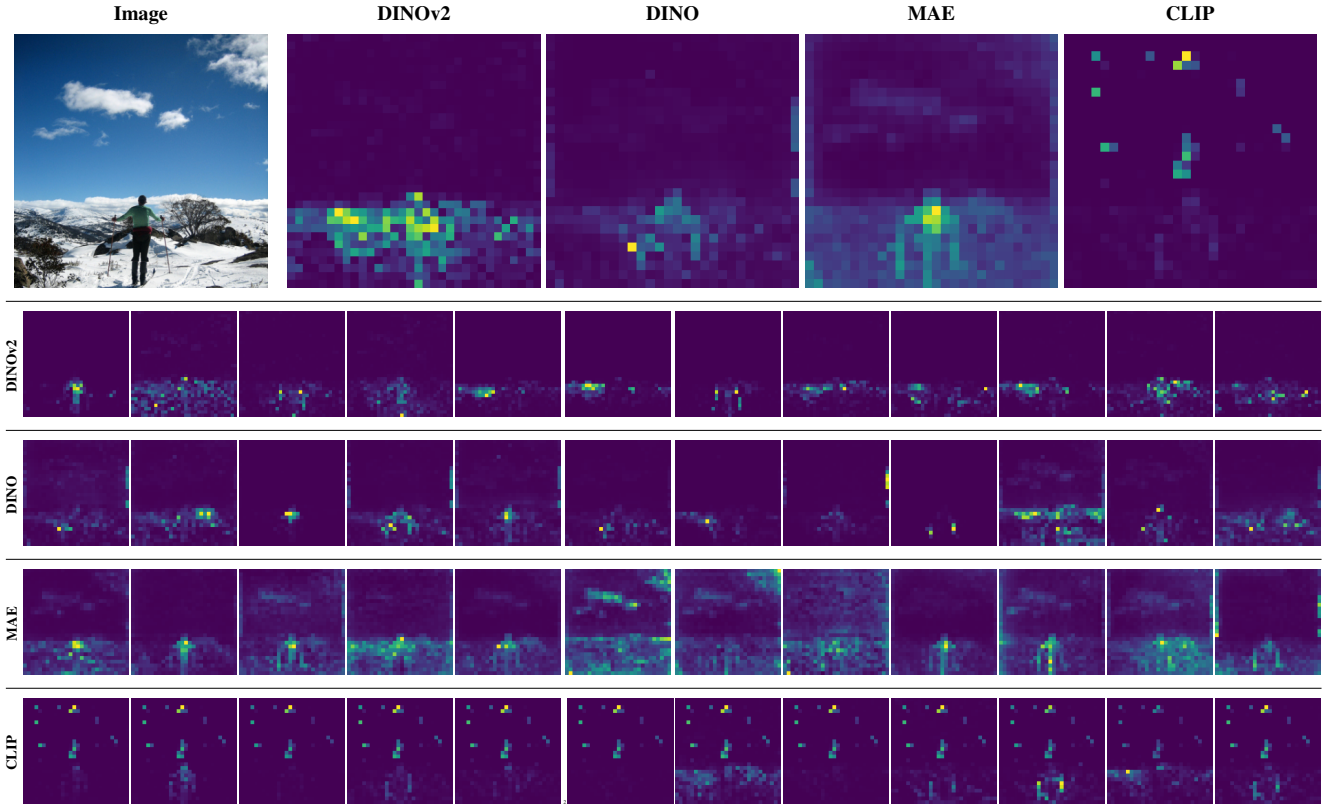

Figure 9. Self-attention activations of different visual backbones (*i.e.*, DINOv2, DINO, MAE, CLIP).

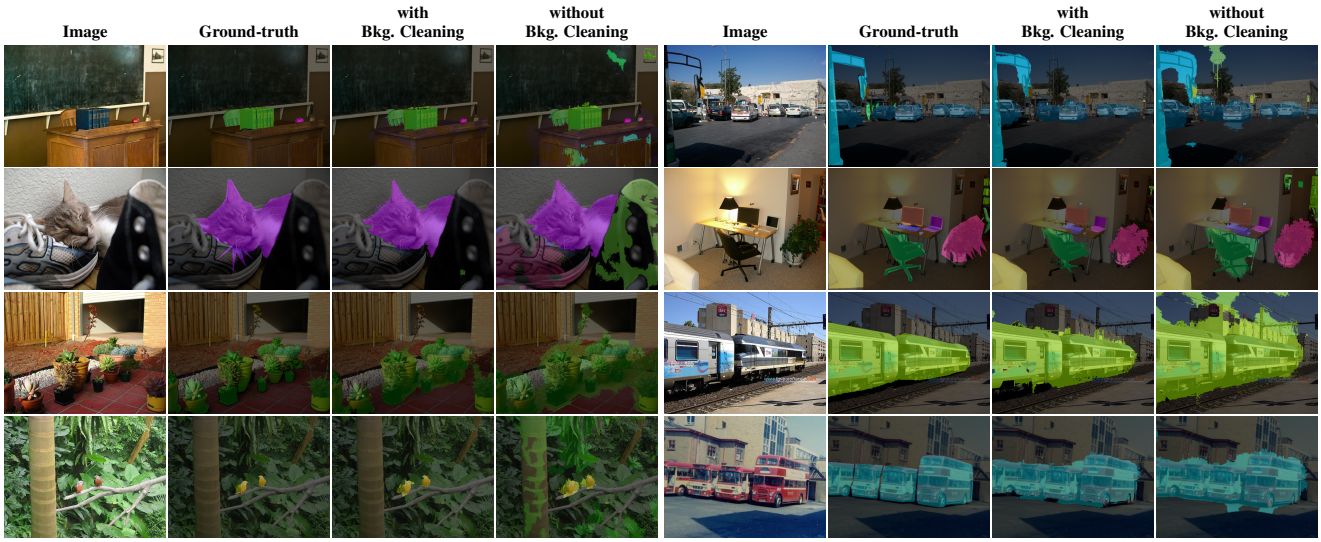

Figure 10. Qualitative results obtained with and without the proposed background cleaning strategy, on COCO Object and Pascal VOC.

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

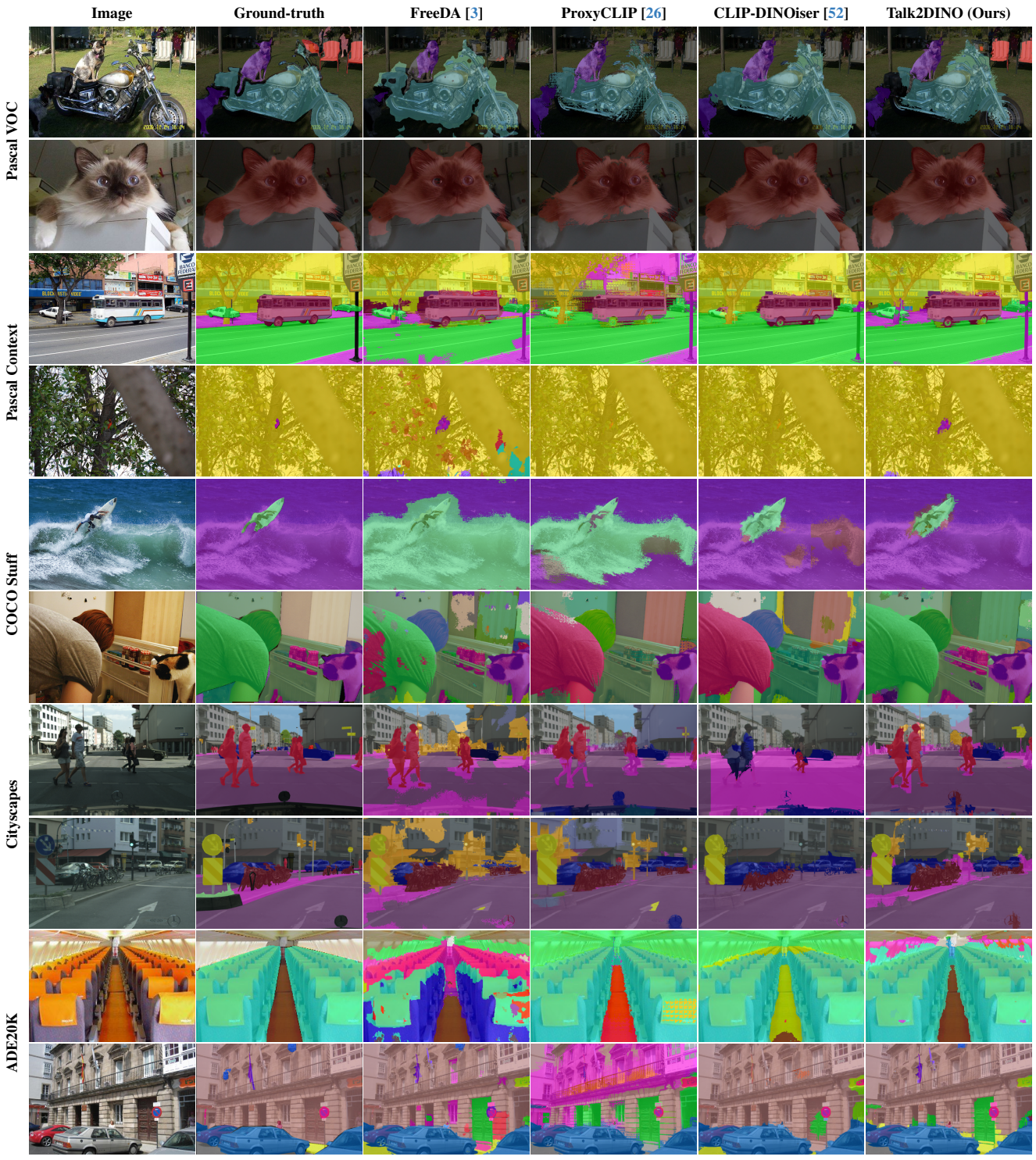

Figure 12. Additional qualitative results of Talk2DINO in comparison with FreeDA [3], ProxyCLIP [26], and CLIP-DINOiser [52].