A. Additional details

Computational cost. In this work, we show that we can obtain good global and local vision-language alignment with minimal additional cost thanks to powerful pre-trained SSL models. This appears to be a more efficient paradigm than training CLIP from scratch. The computational costs for training our models and different CLIP models are reported in Table 7. For completeness' sake, we also include the pretraining cost of the ViT-g DINOv2 vision encoder as well as the cost of distilling this model into a ViT-L. In practice, such additional costs should however be considered amortized over the multiple downstream adaptations of the DINOv2 backbone.

Method	Samples seen	Batch size	GPUs	total GPU.h	GPU arch.
CLIP	12.8B	32768	256	73728	V100
OpenCLIP	12.8B	38400	400	50800	A100 40 GB
MetaCLIP	12.8B	32768	128	92160	V100
EVA-02-CLIP	2B	61000	128	–	A100 40 GB
DINOv2 ViT-g pretraining	–	–	256	22000	A100 80 GB
DINOv2 ViT-L distillation	–	–	–	8000	A100 80 GB
`dino.txt`	3.2B	65536	128	2432	A100 80 GB
`dino.txt` @336	3.2B	65536	256	4096	A100 80 GB

Table 7. **Computational cost of different models in GPU hours.**

ADE20K class names for the error analysis discussion. In Section 4.5, we discuss the failure modes of our zero-shot semantic segmentation method. In particular, we show that class names can be optimized to boost results, instead of using the default ones from each dataset. This is not surprising, the 150 class names of ADE20K were originally chosen to identify each category and were not intended as holistic descriptors for zero-shot segmentation via a vision-language model. In our experiments, we have observed that some class names are too broad, *e.g.*, *building*, or ambiguous, *e.g.*, *throne*, and consequently result in incorrect predictions. In Table 11, we include the optimized class names for ADE20K that improve open-vocabulary segmentation by 2.1 mIoU points, as reported in the discussion about failure modes in Section 4.5. Please note that for all experiments in the main text, we use the original class names to facilitate comparison with previous work.

Example of ambiguous training data. We show in Figure 4 examples of poor image captions of our training data.

- click to enlarge
- ~product.metadata.name~
- Certified pre-owned 2018

Figure 4. **Examples of poor, ambiguous or too generic captions** found in our data pool.

B. Additional ablation studies

Impact of the embedding in segmentation. Table 8 presents open-vocabulary segmentation results on the challenging datasets ADE20K and Cityscapes. We follow the evaluation protocol of TCL [13]. Following only MaskCLIP patch representation (`[value]`) leads to the worst results. Using solely the model's output patch descriptor (`[patch]`) and their corresponding part in the text embedding leads to the best results. This is the setup used in the main paper. We also observe that concatenating the `[CLS]` token to the patch representation hurts the performance *vs.* `[patch]` only, particularly in Cityscapes: we found this to be due to the dominance of the salient visual concept in the `[CLS]`.

Inference embedding	*segmentation* ADE	City.
`[value]` (MCLIP)	7.0	11.7
`[CLS patch]`	19.9	26.2
`[value patch]`	20.0	29.0
`[patch]`	**20.6**	**32.1**

Table 8. **Ablation of the embedding in dense zero-shot segmentation inference.** We show segmentation results with different embeddings to represent a patch, on the datasets ADE20K and Cityscapes. 'MCLIP' corresponds to MaskCLIP [103] strategy, which we also name here `value`.

Impact of the image embedding size at training. We show in Table 9 that the benefit of using the concatenated representation **g** (noted here `[CLS avg]`) when training `dino.txt` does not come from higher dimensionality of the image embedding. To this end, we have conducted an additional experiment in which we project the `[CLS]` token from the dimension of 1024 to 2048 before passing it to the vision blocks. Little impact is observed from this dimensionality change. This additionally shows that the gain (from 30.9 to 34.7) in the retrieval task is largely due to the concatenation of the `[CLS]` token with `[avg]`.

Training embedding	proj	*class.* IN1K	*retr.* COCO
`[CLS]`		78.8	30.2
`[CLS]`	1024 $\rightarrow$ 2048	78.8	30.9
`[CLS avg]`		**79.2**	**34.7**

Table 9. **Analysis of the image embedding size at training time.** Projecting the `[CLS]` embedding to dimension 2048 (second row) yields minimal gain on benchmarks.

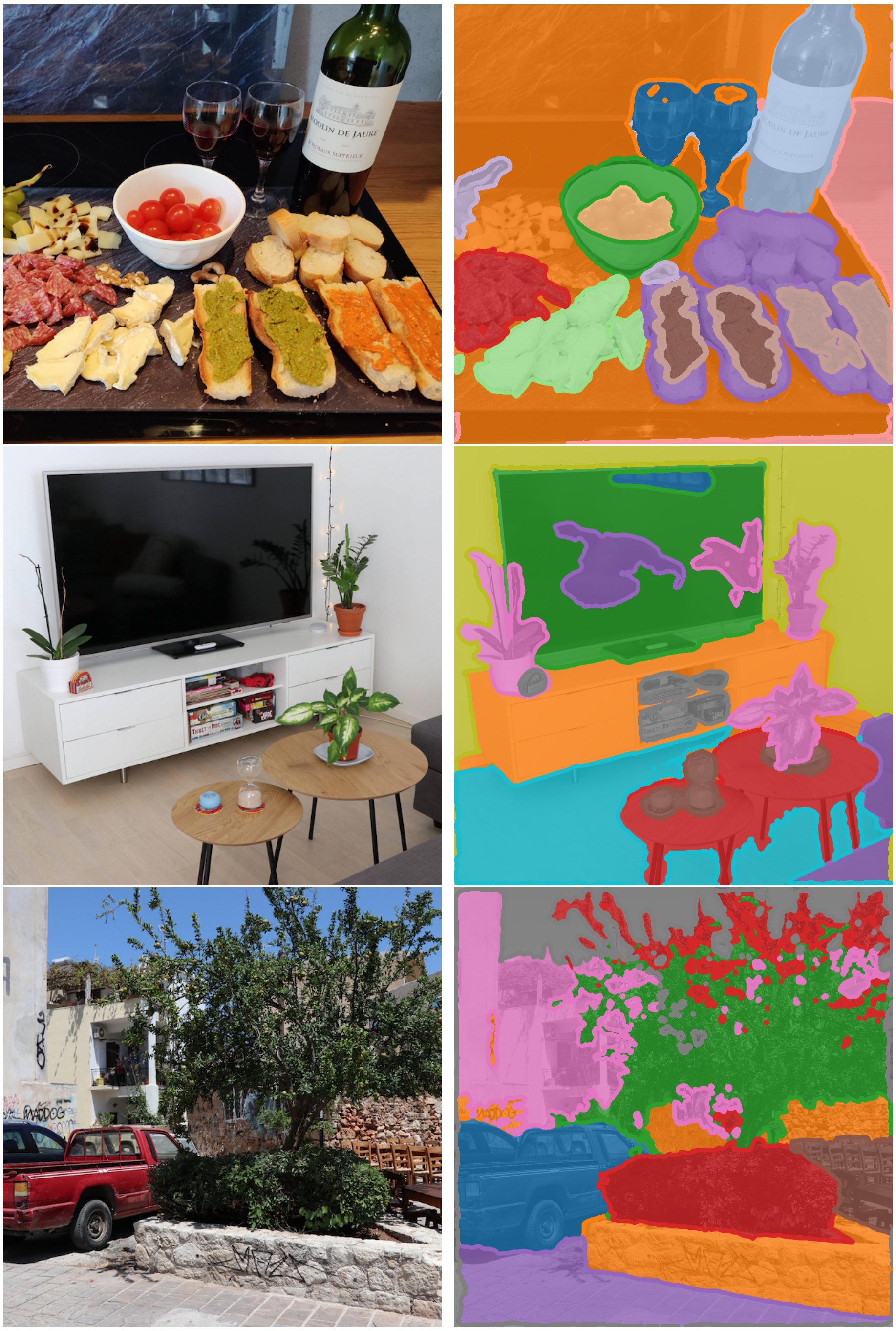
Impact of the trained layer. In this project, we aimed to keep the backbone model as is, with no significant modifications that could alter DINOv2 feature qualities and its performance when considering diverse downstream tasks performed with a single frozen backbone. However, for completeness, we present here additional experiments with no extra block ('none'), or the unfrozen last (two) block(s) (bottom rows). We observe that adding the extra blocks yields the best performance, particularly in segmentation tasks where high-quality localization features from DINOv2 are important. Moreover, unfreezing the last layers give worst results than using the frozen backbone as is, likely due to a degradation of the quality of DINOv2's features.

Trained adapter	IN1K	COCO	ADE
two extra blocks	**81.4**	**45.4**	**20.6**
none	80.9	38.6	17.7
last block	80.7	44.9	17.0
two last blocks	80.6	44.4	13.7

Table 10. **Analysis of the impact of the trained layer.**

C. Additional qualitative results

Open-vocabulary semantic segmentation. Figures 5-6 demonstrate that the segmentation results of `dino.txt` with images and texts in the wild. For each image, we select a small number of descriptive text prompts and run the zero-shot semantic segmentation pipeline described in Section 4.4. Our model is able to segment complex scenes with multiple semantic objects and specific text inputs, *e.g.*, "pesto bruschetta" and "nautical rope".

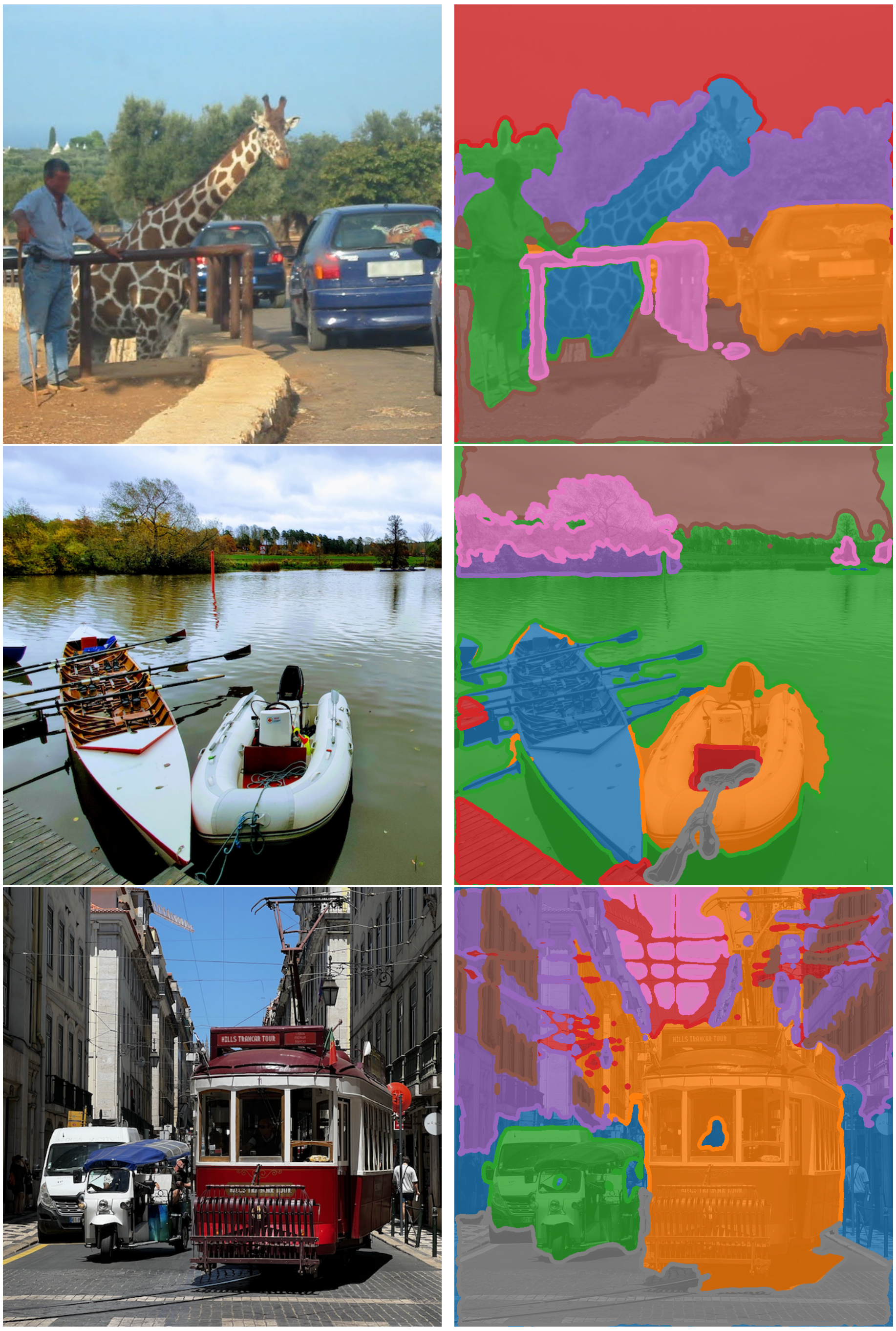


Color	Name
	Wine glass
	Wine bottle
	Stone cutting board
	Cherry tomatoes
	White ceramic bowl
	French cheese
	Salami slices
	Wooden table
	Sliced baguette
	Green grapes
	Pesto bruschetta
	Red pepper spread on bread

Color	Name
	Window
	White cabinet
	Black television screen
	Wooden sofa table
	Gray couch
	Candle
	Potted plant
	Books
	Indoor wall
	Parquet floor

Color	Name
	Red pickup truck
	Stone wall
	Lush tree
	Bush
	Paved road
	Chair
	Facade
	Blue sky

Figure 5. **Open-vocabulary semantic segmentation, part 1/2.** The input resolution is 896×896 pixels.



Color	Name
	Tall giraffe
	Blue automobile
	Tanned man in shirt and pants
	Open sky
	Trees, bushes
	Dirt road, sandy ground
	Wood railing, fence

Color	Name
	Wood rowing canoe
	Inflatable motor boat
	Peaceful lake
	Wooden pier
	Bush
	Blue sky
	Tree
	Nautical rope

Color	Name
	Pedestrian
	Tram
	Car
	Electric wires
	Facade
	Window
	Open sky
	Road, pavement

Figure 6. **Open-vocabulary semantic segmentation, part 2/2.** The input resolution is 896×896 pixels.

Original	Optimized	Original	Optimized
wall	wall	swivel chair	swivel chair
building, edifice	**facade, frontage, frontal**	boat	boat
sky	sky	bar	bar
floor, flooring	**floor**	arcade machine	arcade machine
tree	tree	hovel, hut, hutch, shack, shanty	**hovel**
ceiling	ceiling	bus, autobus, coach, charabanc, double-decker, jitney, motorbus, motorcoach, omnibus, passenger vehicle	**bus**
road, route	**road**	towel	towel
bed	**bed**	light, light source	**skylight, fanlight**
windowpane, window	**windowpane**	truck, motortruck	**truck**
grass	grass	tower	tower
cabinet	cabinet	chandelier, pendant, pendent	**chandelier**
sidewalk, pavement	sidewalk, pavement	awning, sunshade, sunblind	**awning**
person, individual, someone, somebody, mortal, soul	**people**	streetlight, street lamp	**streetlight**
earth, ground	**ground, earth**	booth, cubicle, stall, kiosk	**newsstand**
door, double door	**interior door**	television receiver, television, television set, tv, tv set, idiot box, boob tube, telly, goggle box	**television receiver**
table	table	airplane, aeroplane, plane	**airplane**
mountain, mount	**mountain**	dirt track	dirt track
plant, flora, plant life	**bush**	apparel, wearing apparel, dress, clothes	**clothes closet, clothespress**
curtain, drape, drapery, mantle, pall	**curtain**	pole	pole
chair	chair	land, ground, soil	**land**
car, auto, automobile, machine, motorcar	**car**	bannister, banister, balustrade, balusters, handrail	bannister, banister, balustrade, balusters, handrail
water	water	escalator, moving staircase, moving stairway	**escalator**
painting, picture	**painting**	ottoman, pouf, pouffe, puff, hassock	**footstool, footrest, ottoman, tuffet**
sofa, couch, lounge	sofa, couch, lounge	bottle	bottle
shelf	shelf	buffet, counter, sideboard	**china cabinet, china closet**
house	house	poster, posting, placard, notice, bill, card	**poster**
sea	sea	stage	stage
mirror	mirror	van	van
rug, carpet, carpeting	**rug**	ship	ship
field	field	fountain	fountain
armchair	armchair	conveyer belt, conveyor belt, conveyer, conveyor, transporter	**conveyer belt**
seat	seat	canopy	**baldachin**
fence, fencing	**fence**	washer, automatic washer, washing machine	**washer**
desk	desk	plaything, toy	**plaything**
rock, stone	**rock**	swimming pool, swimming bath, natatorium	**swimming pool**
wardrobe, closet, press	**wardrobe**	stool	stool
lamp	lamp	barrel, cask	**barrel**
bathtub, bathing tub, bath, tub	**bathtub**	basket, handbasket	**basket**
railing, rail	**railing**	waterfall, falls	**waterfall**
cushion	**pillow**	tent, collapsible shelter	**tent**
base, pedestal, stand	**stall, stand, sales booth**	bag	bag
box	box	minibike, motorbike	**motorcycle, bike**
column, pillar	**column**	cradle	**baby bed, baby's bed**
signboard, sign	**signboard**	oven	oven
chest of drawers, chest, bureau, dresser	**chest of drawers**	ball	ball
counter	**reception desk**	food, solid food	**food**
sand	sand	step, stair	**pedestal, plinth, footstall**
sink	sink	tank, storage tank	**tank**
skyscraper	skyscraper	trade name, brand name, brand, marque	**trade name**
fireplace, hearth, open fireplace	fireplace, hearth, open fireplace	microwave, microwave oven	**microwave**
refrigerator, icebox	**refrigerator**	pot, flowerpot	**pot**
grandstand, covered stand	**grandstand**	animal, animate being, beast, brute, creature, fauna	**animal**
path	path	bicycle, bike, wheel, cycle	**bicycle**
stairs, steps	**stairs**	lake	lake
runway	runway	dishwasher, dish washer, dishwashing machine	**dishwasher**
case, display case, showcase, vitrine	case, display case, showcase, vitrine	screen, silver screen, projection screen	**screen**
pool table, billiard table, snooker table	**pool table**	blanket, cover	**blanket**
pillow	**pillow sham**	sculpture	sculpture
screen door, screen	**shower**	hood, exhaust hood	**range hood**
stairway, staircase	**stairway**	sconce	sconce
river	river	vase	vase
bridge, span	**bridge**	traffic light, traffic signal, stoplight	**traffic light**
bookcase	bookcase	tray	tray
blind, screen	**blind**	ashcan, trash can, garbage can, wastebin, ash bin, ash-bin, ashbin, dustbin, trash barrel, trash bin	ashcan, trash can, garbage can, wastebin, ash bin, ash-bin, ashbin, dustbin, trash barrel, trash bin
coffee table, cocktail table	**coffee table**	fan	fan
toilet, can, commode, crapper, pot, potty, stool, throne	**toilet**	pier, wharf, wharfage, dock	**pier**
flower	flower	crt screen	crt screen
book	book	plate	**plate, collection plate**
hill	**hillside**	monitor, monitoring device	**computer screen, computer display**
bench	bench	bulletin board, notice board	**bulletin board**
countertop	countertop	shower	shower
stove, kitchen stove, range, kitchen range, cooking stove	stove, kitchen stove, range, kitchen range, cooking stove	radiator	radiator
palm, palm tree	**cabbage palm, cabbage tree, Livistona australis**	glass, drinking glass	**glass**
kitchen island	kitchen island	clock	clock
computer, computing machine, computing device, data processor, electronic computer, information processing system	**desktop computer**	flag	flag

Table 11. **ADE20K dataset:** original class names *vs*. optimized class names for zero-shot semantic segmentation. Modified class names are highlighted in bold. The new class names have been picked through manual analysis to increase specificity, for example *building* to *facade*, or to remove potential confusion, for example *throne* for *toilet*.