Generalized Decoding for Pixel, Image, and Language Supplementary Material

Xueyan Zou*§, Zi-Yi Dou*♯, Jianwei Yang*‡♠, Zhe Gan†, Linjie Li†, Chunyuan Li‡, Xiyang Dai†, Harkirat Behl‡
Jianfeng Wang†, Lu Yuan†, Nanyun Peng♯, Lijuan Wang†, Yong Jae Lee¶§, Jianfeng Gao¶‡
§ University of Wisconsin-Madison ♯ UCLA ‡ Microsoft Research at Redmond † Microsoft Cloud & AI
*Equal Technical Contribution ¶ Equal Advisory Contribution ♠ Project Lead
{xueyan,yongjaelee}@cs.wisc.edu {zdou,violetpeng}@cs.ucla.edu {jianwyan,jfgao,zhgan,linjli,chunyl,jianfw,luyuan,lijuanw,hbehl,xidai}@microsoft.com

A. Experiment Settings

A.1. Pretraining

During pretraining, we set a minibatch for segmentation to 32 and image-text pairs to 1024. The image resolution is set to 1024 for segmentation and 224 for image-text data respectively. We follow a similar balanced sampling strategy in [11] to ensure the segmentation data are always observed for a consistent number of epochs, regardless of the total number of image-text pairs. Based on this, we pretrain all models for 50 epochs using AdamW [7] as the optimizer.

Also, in the main paper, all the pre-trained models are trained with 50 epochs of COCO data and roughly 45 epochs of 10 million image-text pairs. And 32 GPUs are used for pre-training with 40-50 training hours for Focal-T model. The AdamW optimizer is used in pretraining with the initial learning rate 1e-4. A step-wise scheduler is used to decay the learning rate by 0.1 on the fraction $[0.88889, 0.96296]$ of training steps.

A.2. Finetuning

Image-Text Retrieval. For both COCO and Flickr30k image-text retrieval, we finetune the models for 10 epochs using AdamW as the optimizer. We set the image resolution to 384 and the batch size to 2048. The learning rates are 3e-5 for the X-Decoder part and 3e-6 for the vision and language backbones.

Image Captioning. Similar to image-text retrieval, we finetune the captioning models for 10 epochs using AdamW as the optimizer. We set the image resolution to 480 and the batch size to 256. The learning rates are 2e-5 for the X-Decoder part and 2e-6 for the vision and language backbones. We use beam search during caption generation with the beam size set to 5. We do not use CIDEr optimization for our captioning models.

VQA. For VQA, we add a new classification layer on the top of the model and finetune the models for 10 epochs using AdamW as the optimizer. We set the image resolution to 640 and the batch size to 256. The learning rates are 1e-4 for the X-Decoder part, 1e-5 for the vision and language backbones, and 1e-3 for the VQA classification layer.

Generic Segmentation. For generic segmentation, we finetune the pretrained checkpoint with 24 epochs with start learning rate 1e-4. We decay the learning rate by factor 10 at epoch 21 and 23, respectively. The batch size of ADE20k is 64, and 32 for COCO.

Referring Segmentation. For referring segmentation, we also finetune the pretrained checkpoint with 24 epochs. However, as RefCOCO has been used in pretraining, thus the initial learning rate is 1e-5. It also decays twice at 21 and 23 epochs. We use a batch size of 64 during training. Further, in addition to the normal setting that multiple backbone and language encoder learning rates with 0.1, here we also multiply the transformer encoder learning rate by 0.1.

A.3. Dataset

As mentioned in the main paper, we have exclude the validation sets of Ref-COCOg UMD [13] and COCO Karpathy [12] from COCO2017 [5] in our training setting. This is because the training set of COCO2017 has high overlap with both Ref-COCOg UMD and COCO Karpathy as shown in Fig. 1 below. In addition, we didn't include other validation set of referring segmentation because each split has its own overlap with COCO2017 training set.

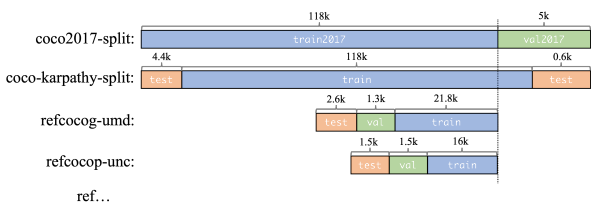


Figure 1. Visualization of dataset overlap. Note that refcocog-umd does also have overlap with refcocop-unc.

B. Open-Vocab Segmentation Benchmark

We propose an open vocabulary segmentation benchmark on 9 datasets with different evaluation metrics. The goal of this benchmark is to provide a comprehensive and standard evaluation protocol for open-vocabulary segmentation on different vocabulary sizes and image domains.

Dataset	Scene	Annotation Format			# Images	# Classes
		Sem	Inst	Pano		
ADE-150	common	✓	✓	✓	2000	150
ADE-847	common	✓	✗	✗	2000	847
Pascal Voc	common	✓	✗	✗	1449	20
Pascal Context-59	common	✓	✗	✗	5105	59
Pascal Context-459	common	✓	✗	✗	5105	459
SUN RGB-D	in-door	✓	✗	✗	5050	37
ScanNet-20	in-door	✓	✗	✓	5436	20
ScanNet-41	in-door	✓	✗	✗	5436	41
Cityscapes	driving	✓	✓	✓	500	19/8/19
BDD	driving	✓	✗	✓	1000	19//40

Table 1. Open-Vocabulary Segmentation Benchmark Statistics.

Table 1 shows the dataset statistics in the benchmark. It supports all generic segmentation tasks including semantic/instance/panoptic segmentation. It covers a variety of scopes ranging from 20 to 847 classes. In addition, the evaluation scene includes common objects, in-door scenes as well as autonomous driving scenarios. To enable a better understanding of the open-vocabulary ability on the training/evaluation datasets. We evaluate the coverage of training datasets captions and evaluation datasets concepts in Fig. 6-12 (we split the caption into single words and phrases to find mappings in categories). The major results of the open-vocabulary segmentation are evaluated in the main paper.

Method	COCO Karpathy				VQAv2
	IR@1	TR@1	CIDEr	BLEU	test-dev
X-Decoder (T)	49.3	66.7	122.3	37.8	70.6
X-Decoder-VL (T)	44.3 $\downarrow_{5.0}$	60.3 $\downarrow_{6.4}$	113.2 $\downarrow_{9.1}$	34.8 $\downarrow_{3.0}$	69.4 $\downarrow_{1.2}$

Table 2. Compare finetuning result between X-Decoder and X-Decoder-VL which merely uses 4M image-text pairs for pretraining.

C. Extra Ablation Studies

C.1. Complementariness between Vision and VL

In our main paper, we observed that the vision-language pretraining objectives including image-text contrastive learning and image captioning have clear benefits to image segmentation, particularly in the zero-shot setting. Here, we further study the role of segmentation objectives in vision-language understanding. To investigate, we remove the segmentation data (COCO panoptic segmentation and referring segmentation) and only pretrain X-Decoder on the four million image-text pairs, denoted by X-Decoder-VL. Afterwards, we transfer the model to downstream VL tasks. As we can see from Table 2, the performance significantly drops across all tasks after removing the segmentation data for pretraining. We suspect that segmentation data can help models to learn more fine-grained visual understanding and consequently benefit vision-language tasks. Along with our findings in the main paper, we conclude that *pixel-level segmentation and vision-language learning are complementary to each other for zero-shot and task-specific transfer*.

C.2. Model Architecture Inspection

In Table. 3, we report the results using three different vision backbone architectures, including Swin [6], FocalNet [10] and DaViT [3]. All models in the first block are with tiny size and trained on the combination of image-label and image-text pairs, following the settings in UniCL [11]. In the second block, all the models are initialized with Florence [14] pre-trained DaVit-d5 model. Through the comparisons, we have the following observations: (1) FocalNet and DaViT achieve better performance than Swin across all metrics. Particularly, FocalNet achieves the best performance on generic and referring segmentation, while DaViT is better on the zero-shot vision-language evaluations; (2) After adding the deformable attention, we can see a boost on supervised segmentation but significant (especially large model) degradation on the open-vocabulary segmentation on ADE20K dataset. Based on these experimental results, we make the design choices as mentioned in our main submission: (1) we remove deformable attention in the favor of open-vocabulary segmentation; (2) we use FocalNet as the tiny vision encoder and train it by ourselves using UniCL, while using DaViT [14] as the base and large vision encoder.

C.3. Open-Vocabulary Generic Segmentation Settings Inspection

In Tab. 4, we study the progressive enrichment of data and training settings as well as the pre-trained model usage. X-Decoder-Seg is the baseline of adding a text encoder to Mask2Former [1] with a learnable language encoder. X-Decoder-Seg$^+$ takes use of caption nouns for Hungarian matching to enrich the vocabulary size. In addition to the main paper, we add row 3 in Tab. 4 to demonstrate the performance of X-Decoder with only coco image text pairs. Comparing 3rd row and 4th row, we find adding extra image-text pairs for pretraining clearl improve open-vocabulary segmentation performance especially when the vocabulary size is large (e.g. ADE-150, CONTEXT-59/459). The way of pretraining vision backbone also matters. Comparing the last two rows side by side, though the backbone model sizes are similar, using ImageNet-21K for pretraining leads to inferior performance on most of the datasets except for CONTEXT-459 which contains most number of categories. These results demonstrate the benefits of using more image-text pairs for pretraining the vision backbone or our X-Decoder.

Method	Backbone	Deformable Attn.	Generic Segmentation						Referring	Retrieval		Captioning	
			COCO			ADE (open)			g-Ref	COCO-Karpathy		COCO-Karpathy	
			PQ	mAP	mIoU	PQ	mAP	mIoU	cIoU	IR@1	TR@1	CIDEr	BLEU
X-Decoder (T)	Swin	✗	50.2	38.8	61.9	17.3	9.4	23.7	55.3	28.0	43.7	79.9	24.2
X-Decoder (T)	Swin	✓	**52.3**	**42.7**	**64.5**	17.0	9.3	22.1	59.1	28.1	43.1	**87.2**	**26.9**
X-Decoder (T)	Focal	✗	51.4	40.5	62.8	**18.8**	**9.8**	**25.0**	**59.8**	30.7	48.5	79.9	24.2
X-Decoder (T)	Davit	✗	51.0	39.7	62.4	17.3	9.4	23.6	58.4	**31.4**	**48.8**	86.8	26.0
X-Decoder (L)	Davit	✗	56.9	46.7	67.7	**21.8**	**13.1**	**29.6**	64.2	44.7	60.3	**111.0**	**32.6**
X-Decoder (L)	Davit	✓	**57.4**	**48.0**	**69.7**	19.1	12.6	26.6	**65.1**	**46.2**	**61.8**	108.2	30.1

Table 3. Model architecture inspection among Swin [6], FocalNet [10] and DaViT [3]. "Deformable Attn." means multi-scale deformable attention [15] that is used in Mask2Former [1]. All numbers are reported in zero-shot manner without any task-specific finetuning, and the row colored in gray corresponds to the architecture used the main paper.

Model	COCO (p/s)			ITP	ADE-150			VOC	PC-59	PC-459	SUN	SCAN-20		SCAN-41	Cityscapes			BDD	
	m	cls	cap		PQ	mAP	mIoU	mIoU	mIoU	mIoU	mIoU	mIoU	PQ	mIoU	mIoU	mAP	PQ	mIoU	PQ
X-Decoder-Seg (T)	✓	✓	✗	✗	13.7	6.3	18.0	89.3	59.3	11.5	16.3	8.6	16.3	6.4	46.6	14.9	30.2	36.9	13.0
X-Decoder-Seg+(T)	✓	✓	✓	✗	15.0	7.8	21.3	93.1	61.7	10.4	28.7	30.7	30.8	17.1	48.2	**16.7**	37.1	40.0	13.5
X-Decoder (T)	✓	✓	✓	✗	16.6	8.3	22.3	94.4	57.6	11.9	33.1	**39.7**	26.4	**21.9**	**51.0**	15.6	35.5	**45.0**	14.4
X-Decoder (T)	✓	✓	✓	✓	**18.8**	**9.8**	**25.0**	**96.2**	**62.9**	**12.3**	**34.5**	37.8	**30.7**	21.7	47.3	16.0	**37.2**	42.4	**16.4**
X-Decoder (L-IN21K)	✓	✓	✓	✓	19.9	11.7	29.6	95.8	54.2	**20.5**	42.4	44.9	29.5	27.4	47.2	18.3	33.3	44.9	15.2
X-Decoder (L)	✓	✓	✓	✓	**21.8**	**13.1**	**29.6**	**97.7**	**64.0**	16.1	**43.0**	**49.5**	**39.5**	**29.7**	**52.0**	**24.9**	**38.1**	**47.2**	**17.8**

Table 4. More open-vocabulary segmentation results. We report the results for our X-Decoder pretrained with COCO segmentation and caption annotations only in 3rd row. Additionally, we compare the model initialized with two different pre-trained large vision backbones, FocalNet-Large and DaViT-d5 trained on ImageNet-21K (row 5) and hundreds of millions of image-text pairs (row 6), respectively.

D. Segmentation In the Wild Benchmark

As shown in the main submission, our X-Decoder exhibits a strong generalization ability to segment images in ten settings of seven datasets from different domains, without any dataset-specific finetuning. Inspired by the object detection in the wild setting proposed in GLIP [4], we resort to more domain-specific datasets on the web to further examine the generality of our model. Specifically, we download 55 instance segmentation datasets from Roboflow [1]. Afterward, we clean the datasets by excluding those containing visually undetectable categories (e.g. Different species of plant) or categories labeled with other languages. In the end, we compile 25 datasets that are suitable for evaluation into *segmentation in the wild* (*SegInW*) benchmark and report instance segmentation mAP. The dataset meta information is listed in Tab. 5, and examplar images are shown in Fig. 3.

On the *SegInW* benchmark, we evaluate zero-shot, few-shot, and fine-tuned segmentation for five models (X-Decoder-Seg$^+$ as baselines, and X-Decoder with different visual backbone) on three different tuning scales. In Fig. 4, we report the zero-shot instance segmentation performance on 25 datasets separately in a descending order. Accordingly, X-Decoder shows reasonably good generalization ability to a wide range of visual and concept domains. Specifically, it achieves higher mAP on common objects like fruits and animals but lower ones on fine-grained datasets like toolkits and rare concepts like rail and brain tumor. In Fig. 5, we further show the line chars for few-shot learning and fully-finetuning, and observe that:

[1] https://roboflow.com/

X-Decoder has privilege on small-scale tuning. As shown in Fig. 5 (a-b), comparing with X-Decoder-Seg$^+$ that only extract noun phrase to increase vocabulary size, X-Decoder performs much better with few-shot/finetune setting. Although X-Decoder (B) and X-Decoder-Seg$^+$ (B) have similar zero-shot performance, the gap increases with the number of images tuned. However, as the number of parameters tuned increased by a large margin Fig. 5 (c), the performance gap between X-Decoder and X-Decoder-Seg$^+$ is shrunk to a small margin.

Zero-Shot gap could be bridged by tuning. X-Decoder (L) and X-Decoder (L-IN21K) are initialized with different pre-trained image backbones. Specifically, X-Decoder (L) is initialized by Florence [14] pre-trained Davit-d5, whereas X-Decoder (L-IN21K) is initialized with FocalNet-L pre-trained on ImageNet-21k [2]. As shown in Fig. 5 (a-c), although the gap between X-Decoder-L and X-Decoder-L-IN21K on the zero-shot setting is relatively large. However, the gap on 5/10/full finetuned settings is much smaller and even cross in some settings.

Tuning class embedding is enough for few-shot settings. As shown in Fig. 5 (e-h), on the smaller scale backbone including (T/B), although tuning the full decoder has a better result, the gap is not obvious on 0-10 shots. And on larger scale models including L/L-IN21K, tuning with class embedding has similar/better results on 0-10 shots.

We show more detailed results in Table 6, Table 7 and Table 8. Similar to main paper, we report the number of parameters tuned in each setting.

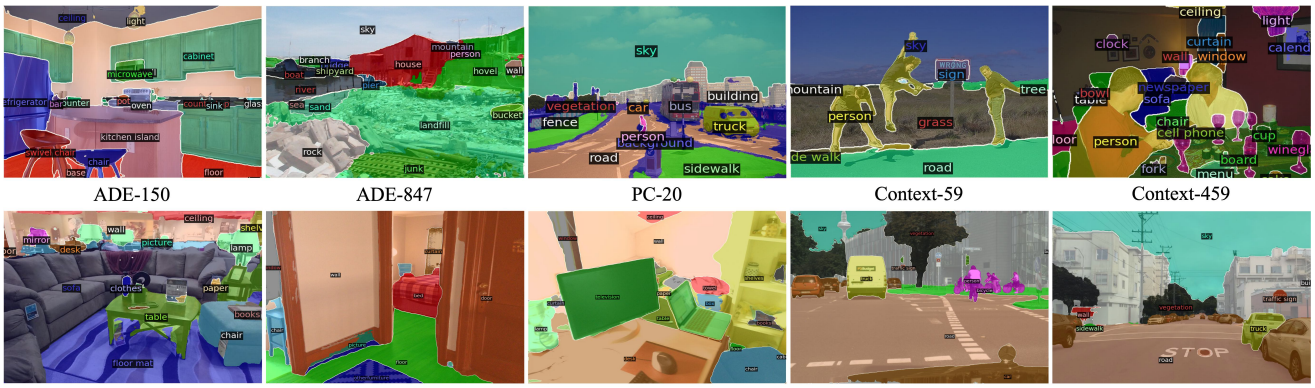


Figure 2. Visualization of zero-shot semantic segmentation on 10 settings of 7 datasets.

Dataset	Categories	# Class	# Images Train	Val	URL
Phones	[phone]	1	25	11	https://universe.roboflow.com/workspace-c4esq/phone-xccez/dataset/1
Elephants	[elephant]	1	883	99	https://universe.roboflow.com/ds/4YwrXdlbFy?key=Q5GY9ITu14
Hand-Metal	[hand, metal]	2	504	65	https://universe.roboflow.com/nk950357-gmail-com/lab-k8hyn
Watermelon	[watermelon]	1	65	23	https://universe.roboflow.com/gnous-b5xq6/my_project-38aqt/dataset/4
House-Parts	[aluminium door, aluminium window, ...]	22	700	201	https://universe.roboflow.com/testcoco/abc-fqun0/dataset/1
HouseHold-Items	[bottle, mouse, perfume, phone]	4	45	3	https://universe.roboflow.com/maths/household-items-sltdd/dataset/2
Strawberry	[R_strawberry, people]	2	971	87	https://universe.roboflow.com/strawberry-25w7z/strawberry_coco_1/dataset/1
Fruits	[apple, lemon, orange, pear, strawberry]	5	120	9	https://universe.roboflow.com/ds/PEVo9xHLFl?key=PXeJGF0D5q
Butterfly-Squirrel	[butterfly, squirrel]	2	951	237	https://universe.roboflow.com/handwashhygeine/nature-3tkys
Hand	[Hand-Segmentation, hand]	2	210	60	https://universe.roboflow.com/rmutsb-xxgii/hand-segmentation-gqzuh/dataset/1
Garbage	[bin, garbage, pavement, road]	4	325	142	https://universe.roboflow.com/project-b1mh9/d2-bj1a0/dataset/1
Chicken	[chicken]	1	19	1	https://universe.roboflow.com/nena-trikic-1jxt3/chickenstf/dataset/1
Rail	[rail]	1	3067	1069	https://universe.roboflow.com/wzk789wzk-gmail-com/rail_dataset/dataset/4
Airplane-Parts	[Airplane, Body, Cockpit, Engine, Wing]	5	39	7	https://universe.roboflow.com/foxehcorp/foxehcorp_airplane_dataset/dataset/4/download
Brain-Tumor	[tumor]	1	236	28	https://universe.roboflow.com/detection-qskiw/segmnetation/dataset/2
Poles	[poles]	1	11	3	https://universe.roboflow.com/ohsee/pole2/dataset/2
Electric-Shaver	[caorau]	1	288	24	https://universe.roboflow.com/fpt-university-1tkhk/caurau
Bottles	[bottle, can, label]	3	357	16	https://universe.roboflow.com/beerup/bottels2/dataset/1
Toolkits	[Allen-key, block, gasket, ...]	8	48	6	https://universe.roboflow.com/mst/mask-2ihnt/dataset/1
Trash	[Aluminium foil, Cigarette, ...]	12	832	92	https://universe.roboflow.com/sara-najafi/trash_segmentation2/dataset/2
Salmon-Fillet	[Salmon_fillet]	1	1991	64	https://universe.roboflow.com/rishik-mishra-rljwe/f1225
Puppies	[puppy]	1	15	3	https://universe.roboflow.com/marcin-bak/puppies-fmoxu/dataset/2
Tablets	[tablets]	1	237	13	https://universe.roboflow.com/detection-qskiw/tablets-instance-segmentation/dataset/1
Cows	[cow]	1	630	60	https://universe.roboflow.com/new-workspace-5abdm/maskrcnn-ofglr/dataset/2
Ginger-Garlic	[garlic, ginger]	2	28	8	https://universe.roboflow.com/george-brown-college-1omrb/ginger-and-garlic-object-segmentation/dataset/1

Table 5. Meta information of *SegInW* benchmark. We list the source links, annotated category names and number of categories for each dataset.

E. Extra Visualization

In this part, we demonstrate the generalization ability to video datasets and flexibility to support task compositions for X-Decoder with more qualitative visualizations.

E.1. Zero-Shot Semantic Segmentation

We visualize zero-shot open vocabulary semantic segmentation on 7 datasets in 10 settings in Fig. 2. The visualization indicates that our model has strong generalization ability on images in different domains as well as categories.

E.2. Zero-Shot Generic Video Segmentation

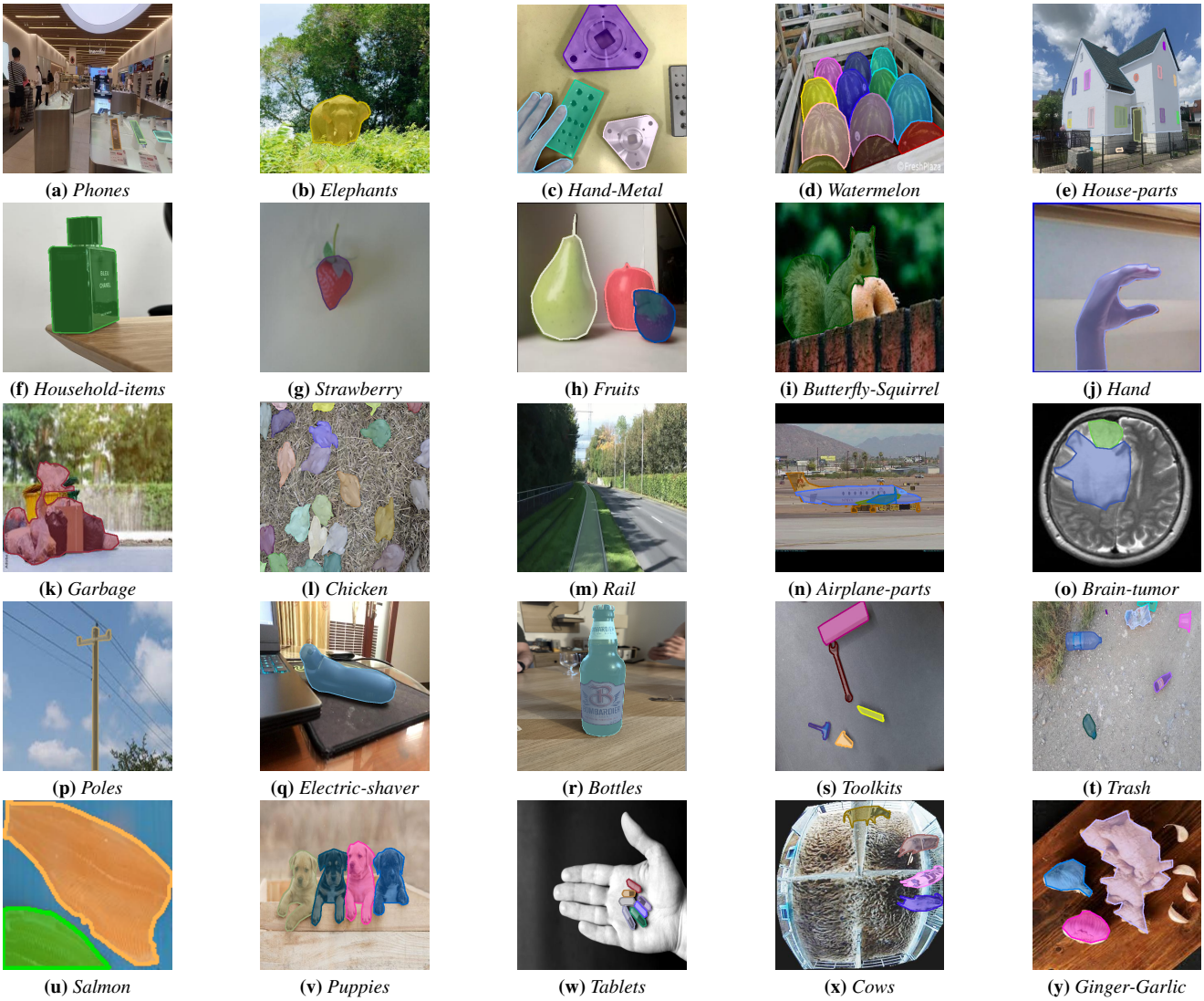
Open-vocabulary generic segmentation is one of the main advantages of X-Decoder. We also apply generic segmentation in a zero-shot manner to the YoutubeVOS [9] dataset. As shown in Fig. 13, our model can be well generalized to video zero-shot generic segmentation and make predictions that are consistent across frames. As a result, our model can be used in video segmentation directly or a good initialization for further finetuning.

E.3. Zero-Shot Referring Video Segmentation

Besides the generic segmentaton on video frames, our X-Decoder can be easily adapted to referring video segmentation as well without any architectural change or finetuning. In Fig. 14, we visualize some examples of referring video segmentation on the YoutubeVOS [9] dataset in a zero-shot manner. We can see that our model can generate rather accurate outputs given various referring phrases. Notably, in addition to the strong segmentation performance for given concepts, the model can also correctly distinguish the spatial locations (*e.g.*, left *v.s.* right in the first row), and object attributes (*e.g.*, a baby gorilla instead of an adult gorilla in the second row) in these unseen videos.

E.4. Zero-Shot Image Captioning

To test the generalization ability of X-Decoder, we also ask the model generate image captions on the YoutubeVOS [9] dataset, which is in a different domain from the image data. As we can see from the examples in Fig. 15, the model can correctly predict the object, activ-



(a) *Phones* **(b)** *Elephants* **(c)** *Hand-Metal* **(d)** *Watermelon* **(e)** *House-parts*

(f) *Household-items* **(g)** *Strawberry* **(h)** *Fruits* **(i)** *Butterfly-Squirrel* **(j)** *Hand*

(k) *Garbage* **(l)** *Chicken* **(m)** *Rail* **(n)** *Airplane-parts* **(o)** *Brain-tumor*

(p) *Poles* **(q)** *Electric-shaver* **(r)** *Bottles* **(s)** *Toolkits* **(t)** *Trash*

(u) *Salmon* **(v)** *Puppies* **(w)** *Tablets* **(x)** *Cows* **(y)** *Ginger-Garlic*

Figure 3. Examplar images and annotations in *SegInW* benchmark. The benchmark covers a diversity of visual domains and concepts in the daily life.

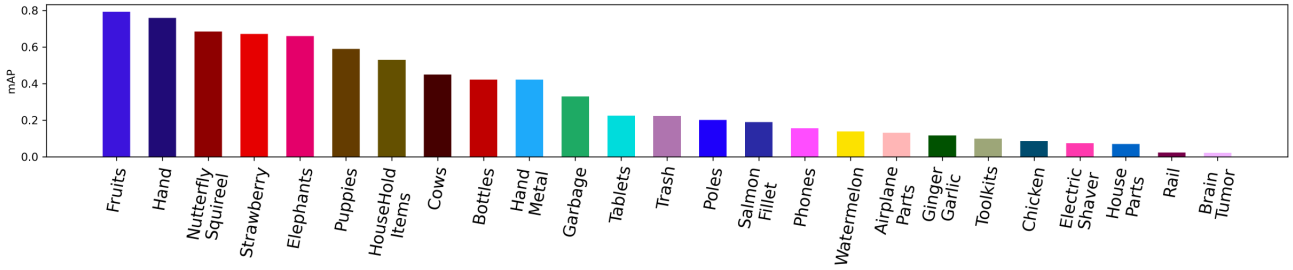
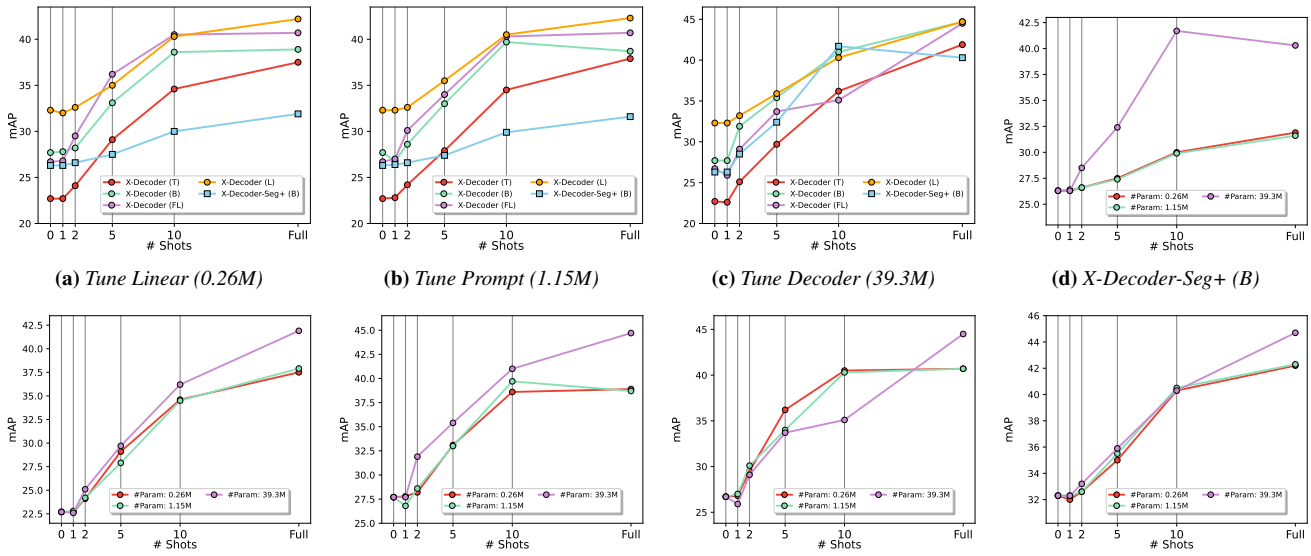


Figure 4. Zero-shot segmentation performance on *SeginW* with X-Decoder-L model. We report the mAP in descending order.

ity, and environment in an image. Interestingly, the captions for the first 6 images sampled from 3 different videos show that our approach can correctly differentiate the movements from similar scenarios (*e.g.*, a man playing vs. a man standing in the first two samples.).

E.5. Zero-Shot Referring Captioning

In compensating for the visualization of the main paper, we add more referring captioning samples in Fig 16. The phrase before ":" is the referring phrase, and the sentence after ":" is the generated caption. The grounding mask of the referring phrase is highlighted in pink. Clearly, our model



(a) *Tune Linear (0.26M)* **(b)** *Tune Prompt (1.15M)* **(c)** *Tune Decoder (39.3M)* **(d)** *X-Decoder-Seg+ (B)*

(e) *X-Decoder (T)* **(f)** *X-Decoder (B)* **(g)** *X-Decoder (L-IN21K)* **(h)** *X-Decoder (L)*

Figure 5. (a-c) Line chart of tuning shot and mAP with different tuning strategies on each backbone architecture. (d-h) Line chart of tuning shot and mAP with different backbone architecture on different number of tuning parameters (strategies).

can simultaneously segments the referred region and generates a region-specific caption. Complementary to regular image captioning systems, such a novel functionality provides a way of interpreting images in a more fine-grained manner. Note that our X-Decoder was never trained to generate such regional captions.

E.6. Zero-Shot Referring Image Editing

Finally, given the high-quality referring segmentation results with X-Decoder, we can effortlessly combine it with off-the-shelf Stable-Diffusion image inpainting model [8] and perform zero-shot referring image editing. As shown in Fig. 17, the model first performs referring segmentation, then the original image and the segmentation mask are fed into the inpainting model to generate the inpainted image. For example, given "change bird to squirrel", it first extracts the bird segment (blue region) from the input image and then replace the segmented region with a generated squirrel. Likewise in other samples, we can see all the generated images look natural and follow the inpainting instructions very well. These impressive plug-and-play results imply a great potential of combining our X-Decoder and advanced generative AI models for fine-grained precise image editing.

F. Discussions

Future Directions. The extensive quantitative and qualitative results have demonstrated the strong performance and generalization ability of our X-Decoder for a variety of vision and vision-language tasks at different granularities. Upon the current X-Decoder design, we see two directions worth future explorations: (1) *Pretrain the whole model in one stage effectively and efficiently*. Currently, the model still requires a separate pretraining for the image and text encoders. However, since our model supports large-scale image-text contrastive learning thanks to the decoupled design, we can easily unify the CLIP-style pretraining with the decoder pretraining in an end-to-end manner. (2) *Unify all level of supervisions*. Due to high annotation costs, the pixel-level segmentation annotations by nature are much less than the region-level box and image-level annotations. It is worth building a more unified learning paradigm to jointly learn from pixel-level, region-level and image-level supervision to attain a more powerful unified model.

Social Impact. This work is mainly focused on the design of a generalized decoder for various vision and vision-language tasks. We have used a pretrained image and text encoder and further pretrained the models on a combination of various datasets and tasks. Since the models are trained on large-scale webly-crawled image-text pairs, the negative impact might arise due to the potential offensive or biased content in the data. To mitigate this issue, we need to have a careful sanity check on the training data and model predictions before deploying it in practical scenarios.

Model	Shot	#Param	Avg	Airplane-Parts	Bottles	Brain-Tumor	Chicken	Cows	Electric-Shaver	Elephants	Fruits	Garbage	Ginger-Garlic	Hand	Hand-Metal	House-Parts	HH.-Items	Butterfly-Squirrel	Phones	Poles	Puppies	Rail	Salmon-Fillet	Strawberry	Tablets	Toolkits	Trash	Watermelon
X-Decoder (T)	0	0.0M	22.7	10.5	19.0	1.1	12.0	12.0	1.2	65.6	66.5	28.7	7.9	0.6	22.4	5.5	50.6	62.1	29.9	3.6	48.9	0.7	15.0	41.6	15.2	9.5	19.3	16.2
X-Decoder (T)	1	39.3M	22.7	10.5±0.0	19.0±0.0	1.1±0.0	12.0±0.0	12.0±0.0	0.8±0.7	65.6±0.0	66.5±0.0	28.7±0.0	7.9±5.8	0.6±0.0	22.4±0.0	5.5±0.0	50.6±4.6	62.1±0.0	29.9±2.3	3.6±0.0	48.9±4.6	0.7±0.0	15.0±1.1	41.6±0.0	15.2±0.0	9.5±0.0	19.3±0.0	16.2±0.0
X-Decoder (T)	3	39.3M	24.1	10.5±0.0	19.0±0.0	1.1±0.0	27.2±26.3	13.4±0.6	1.2±0.0	65.6±0.0	66.9±0.6	28.7±0.0	9.9±2.5	0.6±0.0	21.2±2.2	6.1±0.1	50.6±4.6	62.1±0.0	36.6±7.5	7.6±8.0	49.8±2.5	0.7±0.0	15.0±1.1	41.0±1.0	14.3±1.5	11.5±0.6	19.8±0.7	20.3±5.7
X-Decoder (T)	5	39.3M	29.1	10.0±0.9	30.2±1.2	6.3±3.9	51.8±6.6	18.2±2.7	2.3±1.5	64.9±0.8	67.2±2.1	33.2±1.2	16.9±3.3	30.3±22.8	23.8±6.4	7.1±1.2	50.6±0.1	66.4±1.1	46.4±5.9	14.0±4.8	49.0±0.4	1.5±0.9	15.7±4.0	42.0±0.7	16.1±3.2	12.7±1.6	21.1±1.0	27.9±6.7
X-Decoder (T)	10	39.3M	34.6	10.0±2.0	34.1±6.3	9.7±2.3	59.6±2.4	19.8±5.8	17.4±23.3	64.8±2.1	68.2±7.1	38.1±1.9	17.6±5.9	81.6±4.0	45.6±7.2	8.4±0.9	51.0±0.4	63.9±3.1	41.3±8.4	3.6±0.0	48.4±2.4	1.7±1.5	29.7±7.0	44.4±3.0	24.1±5.7	14.1±2.1	21.6±1.1	44.5±2.4
X-Decoder (T)	3×All	39.3M	37.5	10.5±0.5	35.1±4.4	6.6±1.2	14.5±4.3	30.9±0.9	7.1±0.5	68.3±0.4	70.3±6.9	36.9±0.8	9.9±3.5	37.7±20.1	46.0±3.7	8.5±0.0	50.6±0.0	79.8±0.3	48.7±0.8	13.1±9.4	50.6±0.2	59.6±0.9	54.2±0.5	83.4±10.5	27.4±0.8	13.2±1.0	27.6±0.9	44.5±0.9
X-Decoder-Seg$^+$ (B)	0	0.0M	26.3	13.2	17.2	0.8	33.0	28.6	4.9	67.9	71.1	28.8	5.2	0.0	0.8	6.8	50.6	53.2	18.8	17.9	68.2	0.7	21.1	86.3	5.8	11.5	12.1	31.7
X-Decoder-Seg$^+$ (B)	1	39.3M	26.3	13.2±0.0	17.2±0.0	0.8±0.0	33.0±0.0	28.6±2.3	4.9±0.0	68.0±0.2	71.1±0.0	28.8±0.0	5.2±0.0	0.0±0.0	0.8±0.0	6.8±0.0	50.7±0.0	53.2±0.0	18.8±0.0	17.9±0.0	68.2±0.0	0.7±0.0	21.0±0.0	86.5±0.3	5.8±5.8	11.5±0.0	12.1±1.1	31.7±2.3
X-Decoder-Seg$^+$ (B)	3	39.3M	26.6	13.2±0.0	17.2±0.2	0.9±0.1	33.0±0.0	29.6±0.6	4.4±0.8	67.3±0.5	74.4±4.9	28.6±0.3	6.0±0.8	0.0±0.0	1.1±0.2	7.0±0.2	50.7±0.0	53.1±0.1	23.3±3.3	17.9±0.0	67.1±1.9	0.7±0.0	21.4±0.4	86.5±0.7	5.8±0.0	10.3±1.0	12.2±0.0	32.9±1.0
X-Decoder-Seg$^+$ (B)	5	39.3M	27.5	13.3±0.1	18.6±1.3	1.4±0.1	34.3±3.5	30.6±1.1	5.2±0.7	67.6±2.2	79.5±1.0	28.8±0.1	4.3±0.8	0.0±0.0	1.4±0.3	7.5±0.1	50.7±0.0	54.9±0.9	26.9±4.1	18.8±1.1	66.9±3.0	0.8±0.1	21.6±0.9	85.5±2.1	7.4±0.1	13.0±2.7	12.4±0.0	35.7±1.8
X-Decoder-Seg$^+$ (B)	10	39.3M	30.0	13.5±0.3	18.0±0.2	2.1±0.8	49.7±1.9	35.9±1.8	8.0±3.1	68.2±2.8	79.9±4.0	28.6±0.3	7.3±2.8	0.0±0.0	6.0±1.6	8.1±0.3	50.5±0.0	55.1±2.4	33.6±1.7	17.9±0.0	69.4±3.9	1.6±0.3	24.8±2.5	87.9±0.6	9.0±1.3	14.3±0.6	13.0±0.4	45.7±2.8
X-Decoder-Seg$^+$ (B)	3×All	39.3M	31.9	13.5±0.2	19.8±0.4	2.2±0.1	31.7±2.3	37.7±0.3	4.1±0.3	73.9±0.2	78.8±1.9	29.0±0.0	5.3±0.2	0.0±0.0	3.3±0.6	8.6±0.1	50.7±0.0	63.9±0.0	24.3±1.1	20.1±0.0	65.4±0.0	46.4±1.5	54.0±1.6	89.6±0.0	8.4±0.8	12.7±0.9	13.7±0.1	39.1±1.0
X-Decoder (B)	0	0.0M	27.7	13.0	45.9	0.3	13.6	36.8	4.2	68.0	76.7	30.2	19.4	20.6	18.5	6.7	51.7	53.1	8.9	5.6	55.4	0.8	18.2	81.6	8.0	13.9	27.3	13.0
X-Decoder (B)	1	39.3M	27.8	13.0±0.0	45.9±0.0	0.3±0.0	13.6±1.1	36.8±0.0	4.2±0.0	68.0±0.0	76.7±0.0	30.2±0.0	19.4±0.0	20.6±0.0	18.5±0.0	6.7±5.8	51.7±0.0	53.1±4.6	10.2±2.1	5.6±0.0	55.4±0.0	0.8±0.0	18.2±0.0	81.6±0.0	8.0±0.0	13.9±0.0	27.3±0.0	13.0±0.0
X-Decoder (B)	3	39.3M	28.2	12.6±0.3	41.6±5.2	0.3±0.0	13.6±1.1	37.2±0.6	4.2±0.0	68.0±0.1	77.4±1.1	30.4±0.2	19.8±0.6	26.7±9.6	22.2±6.4	7.2±0.2	51.9±0.3	53.1±4.6	12.2±2.8	5.5±0.4	55.4±0.0	0.3±0.5	19.4±3.4	82.9±1.5	8.7±0.7	10.9±3.1	28.0±0.6	15.3±2.0
X-Decoder (B)	5	39.3M	33.1	12.6±0.4	41.2±0.9	2.2±2.0	27.5±7.5	37.9±0.3	9.8±4.6	67.6±1.2	78.3±1.7	30.4±0.3	33.9±5.8	75.2±14.1	35.5±16.5	8.5±0.4	53.0±1.8	61.9±7.8	12.3±2.7	4.5±0.5	54.7±0.8	0.9±0.1	20.7±0.4	86.7±0.8	8.8±2.5	14.9±6.8	28.9±1.3	19.4±6.8
X-Decoder (B)	10	39.3M	38.6	10.9±1.0	38.3±3.5	1.8±0.9	46.3±4.2	38.6±1.8	24.5±10.3	67.0±4.4	78.4±3.3	30.8±0.4	37.9±2.1	89.1±2.9	63.5±1.0	8.7±0.5	58.4±2.9	72.8±3.0	19.5±7.4	5.6±0.0	57.1±1.1	2.3±0.7	24.6±2.4	87.8±2.8	13.4±5.0	15.0±3.3	32.6±0.4	39.9±3.1
X-Decoder (B)	3×All	39.3M	38.9	12.5±0.0	42.7±0.2	1.0±0.0	14.6±1.4	36.8±0.3	17.4±3.1	71.7±0.2	79.7±0.6	31.5±0.1	29.1±1.6	53.6±0.5	65.8±0.5	9.2±0.0	54.0±0.9	82.3±0.2	17.1±3.2	5.7±1.2	55.4±0.4	48.9±2.7	48.3±1.4	90.3±0.0	18.8±2.5	15.0±0.0	36.3±0.5	33.0±2.1
X-Decoder (L-IN21K)	0	0.0M	26.7	12.3	43.2	0.5	3.5	12.3	18.8	63.9	79.1	24.3	15.6	0.0	20.3	4.9	50.5	58.8	43.4	13.4	57.3	1.3	12.3	74.4	6.9	14.6	20.1	13.5
X-Decoder (L-IN21K)	1	39.3M	26.8	12.3±0.0	43.2±4.6	0.5±0.0	3.5±0.0	13.9±2.8	18.8±0.0	63.9±4.6	79.1±0.0	25.1±1.4	15.6±1.1	0.0±0.0	21.4±1.9	4.9±0.0	50.5±0.0	58.8±0.0	43.4±0.0	14.0±0.9	57.3±0.0	1.3±0.0	12.3±0.0	74.4±0.0	6.9±0.0	14.6±0.0	20.1±0.0	13.5±0.0
X-Decoder (L-IN21K)	3	39.3M	29.5	14.0±0.4	44.9±3.0	0.9±0.7	3.5±1.3	27.0±1.0	21.6±4.9	63.7±0.8	79.1±0.0	24.3±0.0	13.8±4.2	29.6±51.3	16.9±5.8	6.0±0.4	50.6±0.1	59.6±1.3	45.2±0.7	18.4±1.7	57.7±1.4	0.5±0.9	16.7±7.7	83.1±1.8	5.4±3.0	15.3±0.9	23.3±3.1	14.7±2.0
X-Decoder (L-IN21K)	5	39.3M	36.2	12.1±1.6	50.4±4.3	0.4±0.0	31.7±6.9	32.7±0.7	51.9±18.6	64.2±0.7	75.7±3.9	27.8±0.7	22.4±13.4	60.0±33.8	23.9±10.0	7.1±0.1	51.4±1.0	63.0±3.6	42.7±2.9	15.7±5.6	59.7±1.7	1.8±0.1	21.4±9.0	83.7±2.5	16.3±11.6	16.5±0.4	34.0±4.7	37.1±2.7
X-Decoder (L-IN21K)	10	39.3M	40.5	11.8±0.4	52.0±2.3	0.6±0.2	34.1±6.1	34.3±1.1	48.7±16.6	65.3±1.7	80.0±0.9	30.4±11.3	28.0±10.8	91.5±2.8	47.4±29.2	7.0±0.5	54.2±5.1	73.0±6.9	44.6±1.8	13.4±0.0	55.0±5.2	4.6±1.3	24.4±3.6	85.3±1.1	24.7±18.6	20.2±1.3	37.0±1.5	43.8±3.1
X-Decoder (L-IN21K)	3×All	39.3M	40.7	14.1±1.0	53.2±0.7	0.4±0.1	4.5±1.7	36.6±0.4	62.7±4.2	70.3±0.1	80.0±1.2	31.7±0.7	15.6±6.0	22.2±0.5	61.7±0.8	7.8±0.2	51.8±1.0	85.2±0.0	44.1±0.3	12.6±5.9	60.7±0.0	41.7±1.0	48.3±0.6	90.4±0.0	22.9±1.7	18.7±1.2	42.3±0.6	36.8±2.4
X-Decoder (L)	0	0.0M	32.3	13.1	42.1	2.2	8.6	44.9	7.5	66.0	79.2	33.0	11.6	75.9	42.1	7.0	53.0	68.4	15.6	20.1	59.0	2.3	19.0	67.1	22.5	9.9	22.3	13.8
X-Decoder (L)	1	39.3M	32.0	13.1±0.0	42.1±0.0	2.2±0.0	8.6±0.0	44.9±0.0	7.5±0.0	66.0±0.0	79.2±0.0	33.0±0.0	11.6±1.1	75.9±0.0	42.1±0.0	7.0±0.0	53.0±0.0	68.4±0.0	15.6±1.1	20.1±0.0	59.0±0.0	2.3±0.0	19.0±0.0	67.1±0.0	22.5±0.0	9.9±0.0	15.4±13.4	13.8±0.0
X-Decoder (L)	3	39.3M	32.6	13.1±0.0	42.1±0.0	2.2±0.0	12.3±6.4	45.1±0.2	7.5±0.0	66.0±0.0	78.6±0.5	33.3±0.5	11.6±1.1	75.9±0.0	42.1±0.0	7.0±0.0	53.0±0.0	68.4±0.0	17.0±5.2	21.6±1.7	59.0±0.0	2.4±0.1	19.0±0.0	67.1±0.0	23.3±1.4	9.5±0.7	22.3±0.0	13.8±0.0
X-Decoder (L)	5	39.3M	35.0	14.0±0.3	45.3±3.6	4.1±0.4	24.9±11.0	46.1±0.1	11.2±7.1	65.8±0.8	77.9±1.1	33.6±0.5	13.2±1.6	85.1±4.1	43.5±4.1	7.4±0.0	52.9±0.2	69.2±1.4	16.9±8.2	21.6±1.6	58.5±3.1	2.6±0.2	18.4±0.9	81.2±3.9	25.8±4.8	9.7±0.3	24.9±1.8	19.6±2.6
X-Decoder (L)	10	39.3M	40.3	13.3±0.3	45.2±3.5	3.2±1.7	42.3±3.9	45.8±0.1	29.3±3.5	68.3±2.0	76.0±3.1	37.9±1.9	24.4±1.3	93.7±0.4	57.5±1.2	7.9±0.5	52.1±0.3	78.8±1.3	27.0±1.5	20.1±0.0	56.7±4.9	3.3±0.3	17.5±0.9	85.2±1.4	40.1±7.0	8.4±0.6	31.4±0.7	42.0±8.7
X-Decoder (L)	3×All	39.3M	42.2	13.9±0.5	48.4±0.2	7.9±2.8	8.6±0.0	45.3±0.2	20.5±0.2	72.4±0.0	80.5±1.0	36.7±1.1	14.8±1.4	86.7±1.8	63.8±0.3	7.5±0.2	52.8±0.8	83.3±0.1	20.1±1.2	18.1±6.6	57.4±3.0	45.1±0.9	50.2±1.1	92.0±0.1	40.4±1.0	10.4±0.7	36.3±0.6	40.2±0.7

Table 6. SegInW results with tuning on class embedding for different image shots and backbone architectures. (39.3M parameters tuned in the setting.)

Model	Shot	#Param	Avg	Airplane-Parts	Bottles	Brain-Tumor	Chicken	Cows	Electric-Shaver	Elephants	Fruits	Garbage	Ginger-Garlic	Hand	Hand-Metal	House-Parts	HH.-Items	Butterfly-Squirrel	Phones	Poles	Puppies	Rail	Salmon-Fillet	Strawberry	Tablets	Toolkits	Trash	Watermelon
X-Decoder (T)	0	0.0M	22.7	10.5	19.0	1.1	12.0	12.0	1.2	65.6	66.5	28.7	7.9	0.6	22.4	5.5	50.6	62.1	29.9	3.6	48.9	0.7	15.0	41.6	15.2	9.5	19.3	16.2
X-Decoder (T)	1	1.15M	22.8	10.5±0.0	19.0±0.0	1.1±0.0	12.0±0.0	12.0±0.0	1.2±0.0	65.6±0.0	66.5±0.0	28.7±0.0	7.9±5.8	0.6±0.0	22.4±0.0	5.5±0.0	50.6±4.6	62.1±0.0	33.9±4.9	3.6±0.0	48.9±4.6	0.7±0.0	15.0±1.1	41.6±0.0	15.2±0.0	9.5±0.0	19.3±0.0	16.2±0.0
X-Decoder (T)	3	1.15M	24.2	10.1±0.1	22.9±6.7	1.6±0.6	19.2±12.4	13.5±0.4	1.2±0.0	66.0±0.4	66.5±0.0	29.9±2.0	7.9±5.8	0.6±0.0	22.4±0.0	5.8±0.4	50.6±4.6	62.1±0.0	38.1±3.9	7.2±8.4	48.9±4.6	0.9±0.3	15.0±1.1	42.4±1.3	16.5±2.4	10.2±0.6	19.9±1.0	24.1±5.0
X-Decoder (T)	5	1.15M	27.9	10.5±0.3	30.1±1.8	4.1±3.3	36.6±3.4	15.6±1.4	2.2±0.7	66.0±0.4	69.7±2.9	31.6±1.2	7.2±0.0	22.3±25.9	31.8±5.9	7.9±0.5	50.6±0.0	65.4±1.4	43.7±5.4	17.9±1.9	51.6±1.0	1.0±0.3	15.7±1.7	39.4±4.7	17.8±4.2	11.8±1.3	21.4±0.2	24.1±7.8
X-Decoder (T)	10	1.15M	34.5	10.1±0.0	38.4±2.6	4.7±1.1	56.3±7.4	20.7±5.6	2.6±1.5	66.8±2.7	64.0±7.7	36.1±2.1	21.8±5.5	84.9±2.6	41.1±14.3	8.3±0.6	53.1±3.2	68.0±1.6	45.4±5.6	3.6±0.0	49.7±0.8	6.7±2.4	24.3±5.3	43.0±1.5	32.1±5.8	14.6±3.1	22.6±1.7	41.8±12.1
X-Decoder (T)	3×All	1.15M	37.9	10.7±0.1	37.1±3.6	7.7±2.9	13.1±2.0	30.5±0.8	7.3±3.5	68.1±0.5	73.7±1.2	38.0±0.8	10.0±2.0	38.1±20.3	41.5±2.9	8.6±0.1	50.6±0.0	80.0±0.2	45.4±2.1	17.9±1.9	50.7±0.9	59.6±2.0	54.6±1.0	89.7±0.1	28.9±2.1	12.9±0.2	27.6±0.4	43.8±0.5
X-Decoder-Seg$^+$ (B)	0	0.0M	26.3	13.2	17.2	0.8	33.0	28.6	4.9	67.9	71.1	28.8	5.2	0.0	0.8	6.8	50.6	53.2	18.8	17.9	68.2	0.7	21.1	86.3	5.8	11.5	12.1	31.7
X-Decoder-Seg$^+$ (B)	1	1.15M	26.4	13.2±0.0	17.2±0.0	0.8±0.0	33.0±0.0	28.6±2.3	4.9±0.0	67.9±0.0	74.1±5.1	28.8±0.0	5.2±0.0	0.0±0.0	0.8±0.0	6.8±0.0	50.6±0.0	53.2±0.0	18.8±0.0	17.9±0.0	68.2±0.0	0.7±0.0	21.1±0.0	86.3±0.0	5.8±5.8	11.5±0.0	12.1±1.1	31.7±2.3
X-Decoder-Seg$^+$ (B)	3	1.15M	26.6	13.3±0.1	17.2±0.0	1.1±0.3	33.0±0.0	28.9±0.5	4.9±0.0	67.2±1.3	74.4±4.9	28.7±0.0	5.2±0.1	0.0±0.0	1.0±0.0	7.0±0.2	50.6±0.0	53.3±0.0	22.1±3.5	18.3±0.3	66.1±1.8	0.7±0.0	21.4±0.8	86.6±0.0	6.2±0.7	11.1±0.5	12.2±0.2	33.5±0.7
X-Decoder-Seg$^+$ (B)	5	1.15M	27.4	13.5±0.1	17.7±1.0	1.6±0.8	31.5±6.9	32.7±3.0	4.6±0.9	68.9±2.9	79.7±1.5	28.7±0.0	6.7±2.8	0.0±0.0	1.2±0.1	7.6±0.3	50.6±0.0	53.7±0.6	25.6±2.6	18.6±1.3	66.1±1.4	0.8±0.1	21.2±0.3	87.2±0.6	6.9±0.9	10.1±0.8	12.6±0.2	35.5±3.2
X-Decoder-Seg$^+$ (B)	10	1.15M	29.9	13.5±0.5	19.4±0.9	2.5±1.1	45.1±10.9	37.5±2.8	7.2±3.9	69.4±2.9	81.4±0.4	28.8±0.3	8.2±2.3	0.0±0.0	2.1±0.3	8.2±0.1	51.9±1.8	54.6±1.7	32.2±7.7	17.9±0.0	71.7±1.5	1.7±0.1	24.5±0.8	87.5±0.2	10.7±1.3	14.9±1.5	12.5±0.1	42.0±3.8
X-Decoder-Seg$^+$ (B)	3×All	1.15M	31.6	13.5±0.1	19.0±0.5	1.7±0.0	33.0±0.0	37.7±0.3	3.8±0.1	73.7±0.2	75.0±3.8	29.0±0.1	5.0±0.4	0.0±0.0	3.6±0.5	8.6±0.2	50.7±0.0	63.7±0.9	24.7±1.2	19.4±1.2	65.3±0.1	45.7±1.0	53.0±1.0	89.6±0.0	8.4±0.7	13.5±1.3	13.4±0.1	38.2±0.5
X-Decoder (B)	0	0.0M	27.7	13.0	45.9	0.3	13.6	36.8	4.2	68.0	76.7	30.2	19.4	20.6	18.5	6.7	51.7	53.1	8.9	5.6	55.4	0.8	18.2	81.6	8.0	13.9	27.3	13.0
X-Decoder (B)	1	1.15M	26.8	13.0±0.0	45.9±0.0	0.3±0.0	9.0±7.8	36.8±0.0	4.1±0.2	68.0±0.0	76.7±0.0	30.2±0.0	19.4±0.0	20.6±0.0	18.5±0.0	6.7±5.8	51.7±0.0	35.4±30.6	8.9±0.0	4.9±0.5	55.4±0.0	0.8±0.0	18.2±0.0	81.6±0.0	8.0±0.0	13.9±0.0	27.3±0.0	13.0±0.0
X-Decoder (B)	3	1.15M	28.6	12.5±0.1	44.3±2.4	0.2±0.0	21.4±7.2	37.7±0.8	5.1±1.5	67.9±0.1	76.6±0.2	30.3±0.0	24.9±4.9	23.4±2.4	18.5±0.0	7.1±0.6	51.7±0.0	53.1±4.6	10.7±5.4	6.7±0.4	55.1±0.3	0.8±0.0	18.6±2.7	84.2±2.3	8.1±1.5	12.6±2.3	27.3±0.0	14.0±1.8
X-Decoder (B)	5	1.15M	33.0	12.4±0.9	43.7±0.9	0.4±0.3	40.0±4.1	38.1±0.7	9.0±3.7	67.4±0.4	79.9±1.2	30.9±0.8	34.8±4.8	34.4±10.0	51.1±13.7	8.4±0.1	54.2±2.1	65.6±2.2	15.1±5.3	4.5±0.4	55.4±0.0	1.1±0.2	24.4±5.6	87.0±2.3	8.5±0.1	12.5±3.5	28.5±0.4	16.6±2.1
X-Decoder (B)	10	1.15M	39.7	10.3±1.8	36.8±4.1	1.4±0.8	49.3±0.9	38.8±1.3	28.4±2.2	69.2±1.6	76.3±0.4	31.4±0.5	37.2±1.1	92.2±0.2	60.1±6.0	8.8±0.7	55.8±3.8	74.2±1.8	28.7±11.4	5.6±0.0	57.0±2.4	3.2±0.4	30.2±2.9	88.0±1.7	19.8±2.3	17.6±6.0	30.9±1.4	40.4±2.5
X-Decoder (B)	3×All	1.15M	38.7	12.6±0.2	41.7±1.7	1.1±0.1	13.6±1.1	36.5±0.1	23.1±7.9	71.5±0.3	77.3±0.3	31.5±0.2	32.4±6.1	47.3±14.1	65.6±0.5	9.2±0.0	53.8±0.3	82.2±0.1	17.1±1.2	5.5±1.4	55.8±0.7	48.4±3.8	47.9±1.2	90.0±0.3	18.3±0.3	12.8±0.7	36.9±1.0	33.2±0.9
X-Decoder (L-IN21K)	0	0.0M	26.7	12.3	43.2	0.5	3.5	12.3	18.8	63.9	79.1	24.3	15.6	0.0	20.3	4.9	50.5	58.8	43.4	13.4	57.3	1.3	12.3	74.4	6.9	14.6	20.1	13.5
X-Decoder (L-IN21K)	1	1.15M	27.0	12.3±0.0	43.2±4.6	0.5±0.0	3.5±0.0	20.1±6.8	18.8±0.0	63.9±4.6	78.6±0.4	24.3±0.0	15.6±1.1	0.0±0.0	20.3±0.0	4.9±0.0	50.5±0.0	58.8±0.0	43.4±0.0	13.6±0.3	57.3±0.0	1.3±0.0	12.3±0.0	76.4±3.4	7.1±0.2	14.6±0.0	20.1±0.0	13.5±0.0
X-Decoder (L-IN21K)	3	1.15M	30.1	12.7±0.8	45.8±4.5	0.6±0.1	15.4±9.0	28.6±1.9	38.0±18.2	64.1±0.2	78.2±1.5	24.3±0.0	15.6±1.1	26.3±23.1	20.9±3.3	5.6±0.4	50.5±0.0	60.8±3.5	42.2±2.3	12.2±3.1	57.3±0.0	1.5±0.3	17.8±9.5	77.8±5.7	7.0±4.4	15.2±1.0	13.4±11.6	19.3±9.9
X-Decoder (L-IN21K)	5	1.15M	34.0	13.8±1.4	48.2±3.1	0.7±0.3	26.3±6.2	28.7±3.9	33.3±25.4	62.8±0.2	79.0±1.9	29.3±2.0	16.4±10.1	60.1±52.0	34.5±22.5	7.2±0.4	52.2±2.5	63.5±1.7	37.0±10.8	7.6±6.2	59.8±0.8	2.3±0.2	21.8±3.5	85.4±1.1	5.7±2.3	16.0±2.1	27.7±10.8	29.8±8.7
X-Decoder (L-IN21K)	10	1.15M	40.3	12.5±1.4	46.5±7.0	0.8±0.5	36.2±3.6	31.9±3.9	44.7±29.6	63.1±1.6	79.8±1.0	33.2±5.5	34.3±3.6	88.9±3.1	62.2±0.8	7.2±1.4	53.2±2.2	76.0±2.1	27.2±21.0	13.4±0.0	58.8±3.7	2.4±0.1	23.3±4.1	88.0±1.4	25.9±12.8	17.6±2.7	34.3±6.9	43.9±1.3
X-Decoder (L-IN21K)	3×All	1.15M	40.7	14.5±0.5	53.3±0.4	0.5±0.1	4.1±1.0	36.8±0.0	64.3±0.2	70.7±0.4	80.7±1.1	32.1±0.1	13.2±5.6	20.4±1.9	61.5±0.7	7.9±0.1	51.6±0.7	84.8±0.0	43.2±1.1	13.5±6.0	60.5±0.3	42.9±2.7	48.8±0.9	90.4±0.3	24.5±3.6	19.2±0.8	41.6±0.3	35.7±1.2
X-Decoder (L)	0	0.0M	32.3	13.1	42.1	2.2	8.6	44.9	7.5	66.0	79.2	33.0	11.6	75.9	42.1	7.0	53.0	68.4	15.6	20.1	59.0	2.3	19.0	67.1	22.5	9.9	22.3	13.8
X-Decoder (L)	1	1.15M	32.3	13.1±0.0	42.1±0.0	2.2±0.0	8.6±0.0	44.9±0.0	7.5±0.0	66.0±0.0	79.2±0.0	33.0±0.0	11.6±1.1	75.9±0.0	42.1±0.0	7.0±0.0	53.0±0.0	68.4±0.0	15.6±1.1	20.1±0.0	59.0±0.0	2.3±0.0	19.0±0.0	67.1±0.0	22.5±0.0	9.9±0.0	22.3±0.0	13.8±0.0
X-Decoder (L)	3	1.15M	32.6	13.3±0.3	41.4±1.2	2.2±0.0	8.6±0.0	45.5±0.4	7.5±0.0	66.4±0.6	79.2±0.0	33.0±0.0	11.6±1.1	75.9±0.0	42.1±0.0	7.1±0.2	53.0±0.0	68.4±0.0	18.0±3.3	20.7±0.4	59.0±0.0	2.3±0.0	19.0±0.0	67.1±0.0	22.0±0.8	9.9±0.0	22.9±0.9	17.1±5.6
X-Decoder (L)	5	1.15M	35.5	13.7±0.5	46.9±4.1	4.0±1.6	33.2±1.0	45.7±0.6	12.1±4.8	65.9±1.6	77.6±0.7	32.9±0.5	20.3±10.3	75.9±0.0	41.4±1.2	6.9±0.5	53.0±0.2	70.3±2.0	20.2±1.5	22.4±1.9	59.1±0.9	3.1±0.4	16.8±2.1	80.1±2.3	27.7±4.3	9.4±0.8	23.9±0.4	23.6±8.2
X-Decoder (L)	10	1.15M	40.5	13.6±0.1	45.4±1.8	4.5±2.4	44.9±0.4	46.0±1.1	35.2±10.8	66.7±3.8	78.6±1.7	39.2±2.6	20.1±5.2	94.3±0.3	59.9±3.4	7.2±0.2	52.0±0.7	77.8±1.3	24.5±2.8	20.1±0.0	56.7±1.3	3.2±0.6	19.7±2.5	86.9±0.4	38.8±4.5	8.8±3.2	30.7±1.2	36.5±4.5
X-Decoder (L)	3×All	1.15M	42.3	13.4±0.0	48.8±0.7	6.3±1.2	8.6±0.0	45.1±0.0	20.5±1.0	72.1±0.1	79.3±0.9	36.9±0.9	12.8±1.1	88.5±2.5	63.1±1.9	7.6±0.0	52.8±0.8	83.6±0.2	22.1±1.1	21.7±1.9	59.2±0.8	43.7±2.7	50.0±1.0	91.7±0.0	40.9±1.4	9.8±0.3	36.6±0.3	40.7±1.7

Table 7. SegInW results with tuning on class & mask embeddings and latent queries for different image shots and backbone architectures. (1.15M parameters tuned in the setting.)

Model	Shot	#Param	Avg	Airplane-Parts	Bottles	Brain-Tumor	Chicken	Cows	Electric-Shaver	Elephants	Fruits	Garbage	Ginger-Garlic	Hand	Hand-Metal	House-Parts	HH.-Items	Butterfly-Squirrel	Phones	Poles	Puppies	Rail	Salmon-Fillet	Strawberry	Tablets	Toolkits	Trash	Watermelon
X-Decoder (T)	0	0.0M	22.7	10.5	19.0	1.1	12.0	12.0	1.2	65.6	66.5	28.7	7.9	0.6	22.4	5.5	50.6	62.1	29.9	3.6	48.9	0.7	15.0	41.6	15.2	9.5	19.3	16.2
X-Decoder (T)	1	0.26M	22.6	10.5±0.0	19.0±0.0	1.1±0.0	12.0±0.0	12.0±0.0	1.2±0.0	65.6±0.0	66.5±0.0	28.7±0.0	7.9±5.8	0.6±0.0	22.4±0.0	5.5±0.0	50.6±4.6	62.1±0.0	28.5±2.1	3.6±0.0	48.9±4.6	0.7±0.0	15.0±1.1	41.6±0.0	15.2±0.0	9.5±0.0	19.3±0.0	16.2±0.0
X-Decoder (T)	3	0.26M	25.1	10.4±0.1	20.2±2.1	3.4±4.0	42.7±17.8	14.4±2.3	1.2±0.0	66.1±0.8	65.5±3.1	28.8±0.2	11.6±6.3	0.6±0.0	22.4±0.0	7.4±1.1	50.6±0.0	62.1±0.0	33.4±14.3	9.6±6.6	49.3±0.6	0.7±0.0	15.0±1.1	41.6±0.0	15.3±3.2	11.8±2.5	20.3±1.7	20.5±3.7
X-Decoder (T)	5	0.26M	29.7	10.5±0.7	33.7±3.6	7.8±2.6	33.0±16.3	15.2±2.6	14.2±12.7	65.5±1.7	65.5±8.8	34.7±3.1	16.7±1.6	51.0±18.1	30.3±4.6	7.9±0.6	51.1±0.8	63.6±4.1	46.0±5.6	13.8±11.0	49.5±0.8	0.5±0.4	18.4±8.5	34.6±2.8	19.7±3.7	13.2±1.7	18.9±1.4	26.8±6.4
X-Decoder (T)	10	0.26M	36.2	10.9±1.6	35.2±5.5	6.2±2.6	61.8±3.1	19.8±5.7	46.2±6.8	66.0±2.7	63.2±10.3	34.9±3.8	19.5±13.6	92.0±1.0	46.1±15.0	10.6±1.2	56.3±3.3	67.9±2.1	33.3±12.5	3.6±0.0	45.8±3.5	6.3±0.6	22.8±8.7	52.4±21.2	25.7±10.4	16.4±6.3	21.2±0.9	39.9±2.1
X-Decoder (T)	3×All	0.26M	41.9	10.7±0.4	42.8±2.2	8.7±0.7	13.6±2.8	30.5±3.5	27.5±11.0	69.0±0.8	70.8±2.8	38.5±0.0	10.3±1.6	74.0±1.2	61.0±2.0	13.9±0.3	50.7±0.0	81.3±0.8	44.2±3.0	20.1±0.0	50.3±0.6	62.6±2.6	55.0±0.6	90.8±0.1	28.0±1.9	18.4±2.7	27.3±0.9	45.3±1.7
X-Decoder-Seg$^+$ (B)	0	0.0M	26.3	13.2	17.2	0.8	33.0	28.6	4.9	67.9	71.1	28.8	5.2	0.0	0.8	6.8	50.6	53.2	18.8	17.9	68.2	0.7	21.1	86.3	5.8	11.5	12.1	31.7
X-Decoder-Seg$^+$ (B)	1	0.26M	26.3	13.2±0.0	17.2±0.0	0.8±0.0	33.0±0.0	28.6±0.0	4.9±0.0	67.9±0.0	71.1±0.0	28.8±0.0	5.2±0.0	0.0±0.0	1.5±1.2	6.8±5.8	50.6±0.0	53.2±0.0	18.8±0.0	17.9±0.0	67.3±1.6	0.7±0.0	21.1±0.0	86.3±0.0	5.8±5.8	11.5±0.0	12.1±1.1	31.8±0.0
X-Decoder-Seg$^+$ (B)	3	0.26M	28.5	13.2±0.3	17.2±0.0	1.4±0.5	33.0±0.0	31.2±3.4	9.5±6.7	67.9±0.0	74.1±5.2	28.8±0.0	7.0±3.1	1.3±0.7	12.4±16.4	7.3±0.3	52.6±3.4	55.8±2.5	31.3±7.3	17.6±2.5	68.2±0.0	0.7±0.0	21.9±3.6	86.9±0.9	9.7±0.7	11.9±0.6	12.0±0.7	39.0±7.2
X-Decoder-Seg$^+$ (B)	5	0.26M	32.4	13.7±0.0	20.7±0.9	2.3±0.5	27.1±14.4	30.2±0.6	41.7±27.2	67.3±1.7	77.5±0.2	28.0±1.0	17.8±7.7	3.2±1.8	28.3±11.3	7.2±0.4	50.8±0.1	58.5±4.3	39.6±12.5	20.4±0.3	69.6±2.3	1.0±0.4	31.4±5.7	86.6±1.5	10.9±3.5	14.1±1.3	13.0±1.0	48.3±5.4
X-Decoder-Seg$^+$ (B)	10	0.26M	41.7	14.2±1.7	24.9±4.3	5.5±0.1	66.5±2.7	36.5±0.4	68.0±5.8	69.3±2.6	69.8±9.4	28.5±0.5	24.9±6.0	94.2±0.6	46.4±7.1	9.5±0.0	50.5±0.1	70.3±3.8	52.4±2.2	17.9±0.0	59.5±5.2	16.0±12.1	37.6±6.1	87.6±0.5	13.4±0.3	13.1±3.3	11.8±0.7	51.8±3.2
X-Decoder-Seg$^+$ (B)	3×All	0.26M	40.3	13.5±0.2	21.8±1.2	3.3±0.5	33.0±0.0	40.9±0.9	68.1±6.8	73.2±0.0	73.0±5.4	30.3±0.6	11.6±3.2	26.1±14.4	49.2±2.1	11.4±0.0	50.8±0.1	81.6±0.8	35.6±5.0	20.7±0.9	64.7±1.2	59.2±0.7	51.7±0.8	90.5±0.8	12.8±1.2	15.5±0.3	17.5±0.5	50.6±0.4
X-Decoder (B)	0	0.0M	27.7	13.0	45.9	0.3	13.6	36.8	4.2	68.0	76.7	30.2	19.4	20.6	18.5	6.7	51.7	53.1	8.9	5.6	55.4	0.8	18.2	81.6	8.0	13.9	27.3	13.0
X-Decoder (B)	1	0.26M	27.7	13.0±0.0	45.9±0.0	0.3±0.0	13.6±1.1	36.8±0.0	4.2±0.0	68.0±0.0	76.7±0.0	30.2±0.0	19.4±0.0	20.6±0.0	18.5±0.0	6.7±5.8	51.7±0.0	53.0±4.6	9.9±1.6	5.6±0.0	55.4±0.0	0.8±0.0	18.2±0.0	81.6±0.0	8.0±0.0	13.9±0.0	27.3±0.0	13.0±0.0
X-Decoder (B)	3	0.26M	31.9	13.1±0.4	47.4±3.9	0.3±0.1	18.9±9.2	38.1±2.2	9.4±5.4	69.6±1.4	76.8±0.0	30.9±1.1	19.4±0.0	84.9±1.7	32.2±23.7	8.2±1.3	52.8±1.9	53.0±4.6	10.9±2.0	5.3±0.7	55.4±0.0	0.8±0.0	18.5±1.1	81.6±0.0	8.5±1.3	15.5±2.6	27.4±0.2	17.5±1.2
X-Decoder (B)	5	0.26M	35.4	12.6±0.5	48.7±1.8	1.0±0.8	19.0±16.0	37.1±1.2	21.3±21.3	67.9±1.8	79.4±3.6	32.2±1.7	32.2±3.0	82.8±7.8	63.4±5.0	8.6±0.2	53.7±2.7	65.6±10.9	17.8±6.9	5.0±1.4	54.2±2.1	1.2±0.4	21.0±1.3	81.6±10.0	8.6±3.3	18.2±6.6	30.5±0.6	19.2±3.2
X-Decoder (B)	10	0.26M	41.0	13.7±0.8	42.4±2.8	4.0±2.2	50.8±3.8	40.3±0.8	70.9±6.9	68.8±2.9	78.1±0.6	30.8±6.4	40.4±9.1	76.6±23.9	63.0±3.6	10.7±0.6	60.6±0.6	69.8±1.7	21.0±20.7	5.6±0.0	55.4±1.4	4.4±2.1	27.8±5.4	88.3±2.1	22.4±8.9	14.2±2.7	31.1±2.3	31.4±9.1
X-Decoder (B)	3×All	0.26M	44.7	13.0±0.0	43.8±2.8	3.3±0.0	15.4±3.1	36.5±1.2	69.3±9.3	72.2±0.5	79.6±1.1	34.0±0.7	38.9±1.0	89.4±3.3	74.8±0.9	14.1±0.2	57.9±1.9	84.0±0.4	17.8±3.1	5.1±0.4	55.9±0.3	57.5±1.8	48.4±0.4	90.0±0.1	18.4±0.2	21.0±4.0	38.3±0.5	37.0±0.5
X-Decoder (L-IN21K)	0	0.0M	26.7	12.3	43.2	0.5	3.5	12.3	18.8	63.9	79.1	24.3	15.6	0.0	20.3	4.9	50.5	58.8	43.4	13.4	57.3	1.3	12.3	74.4	6.9	14.6	20.1	13.5
X-Decoder (L-IN21K)	1	0.26M	25.9	12.3±0.0	28.8±24.9	0.5±0.0	3.5±0.0	12.3±1.1	18.8±0.0	63.9±4.6	79.1±0.0	24.3±0.0	15.6±1.1	0.0±0.0	20.3±0.0	4.9±0.0	50.5±0.0	58.8±0.0	43.4±0.0	11.2±3.8	57.3±0.0	1.3±0.0	12.3±0.0	74.4±0.0	5.4±2.5	14.6±0.0	20.1±0.0	13.5±0.0
X-Decoder (L-IN21K)	3	0.26M	29.1	12.6±2.1	44.7±2.6	1.1±0.9	8.0±11.0	15.7±3.0	32.7±24.0	63.9±4.6	76.7±4.4	24.5±0.4	15.6±1.1	30.2±52.4	16.9±5.9	6.8±2.0	51.0±0.7	61.1±4.0	43.0±3.8	14.6±5.8	57.5±3.1	1.4±0.1	12.3±0.0	74.0±0.7	4.4±2.7	14.8±0.3	20.8±1.2	21.0±10.7
X-Decoder (L-IN21K)	5	0.26M	33.7	13.9±1.0	46.4±5.9	2.1±2.0	9.5±8.5	31.4±1.3	52.6±12.3	64.1±0.6	78.0±6.2	32.9±1.2	19.2±8.2	71.0±24.3	26.3±20.1	7.5±0.8	54.9±4.7	66.6±1.7	32.6±9.3	10.3±9.2	59.0±1.6	1.3±0.6	15.5±8.3	80.1±5.5	3.0±3.6	14.0±5.0	26.6±9.2	22.3±8.5
X-Decoder (L-IN21K)	10	0.26M	35.1	10.4±1.5	39.1±7.2	4.4±1.1	31.7±8.5	24.7±11.7	55.8±6.0	61.4±6.1	73.9±7.1	28.6±1.8	17.5±10.2	85.4±7.0	40.8±28.3	6.4±2.1	58.4±2.9	54.2±4.3	32.2±19.3	13.4±0.0	40.2±13.5	2.2±2.1	20.8±14.8	81.0±2.0	17.9±15.8	17.6±3.1	26.4±4.4	31.9±9.4
X-Decoder (L-IN21K)	3×All	0.26M	44.5	12.1±0.8	57.0±0.4	1.5±0.5	4.9±2.3	41.4±1.0	74.7±2.9	70.3±0.4	79.1±2.6	36.6±0.7	23.6±3.8	54.6±2.3	70.0±1.6	12.7±0.1	60.1±0.0	86.1±0.3	43.1±2.1	5.2±2.2	59.7±0.9	46.6±0.8	52.0±0.4	91.0±0.0	23.0±5.5	22.7±1.7	43.8±0.5	40.5±1.3
X-Decoder (L)	0	0.0M	32.3	13.1	42.1	2.2	8.6	44.9	7.5	66.0	79.2	33.0	11.6	75.9	42.1	7.0	53.0	68.4	15.6	20.1	59.0	2.3	19.0	67.1	22.5	9.9	22.3	13.8
X-Decoder (L)	1	0.26M	32.3	13.1±0.0	42.1±0.0	2.2±0.0	8.6±0.0	44.9±0.0	7.5±0.0	66.0±0.0	79.2±0.0	33.0±0.0	11.6±1.1	75.9±0.0	42.1±0.0	7.0±0.0	53.0±0.0	68.4±0.0	15.6±1.1	20.1±0.0	59.0±0.0	2.3±0.0	19.0±0.0	67.1±0.0	22.5±2.3	9.9±0.0	22.3±0.0	13.8±0.0
X-Decoder (L)	3	0.26M	33.2	12.9±0.4	45.9±6.4	1.8±0.6	8.6±0.0	44.9±0.0	7.5±0.0	66.0±0.0	79.2±0.0	33.0±0.0	13.2±2.7	75.9±0.0	42.1±0.0	7.2±0.3	53.0±0.0	68.4±0.0	18.1±0.9	22.4±1.9	59.0±0.0	2.3±0.0	19.8±1.3	67.1±0.0	26.0±6.0	9.6±0.4	25.8±6.0	18.3±5.1
X-Decoder (L)	5	0.26M	35.9	12.5±0.5	44.9±2.2	2.4±2.5	28.4±6.8	44.9±1.0	15.7±1.6	67.1±2.0	77.1±0.2	36.3±0.7	9.8±8.7	93.1±0.9	45.6±7.5	7.6±1.0	53.0±0.8	71.3±3.1	19.4±3.5	22.5±1.7	55.8±2.2	2.4±0.5	12.0±2.1	78.4±6.0	30.1±6.3	10.0±1.3	30.1±1.7	25.9±2.8
X-Decoder (L)	10	0.26M	40.3	14.0±1.5	33.7±14.6	4.4±5.9	41.2±3.0	44.6±1.7	73.0±2.2	68.8±5.5	79.4±1.8	39.2±3.4	17.5±4.7	93.9±0.6	53.9±3.1	8.8±3.0	52.5±2.5	77.3±0.9	24.0±7.0	20.1±0.0	55.3±0.7	3.0±1.6	15.0±6.6	72.7±16.1	39.1±7.8	9.0±5.5	32.1±2.4	32.5±9.4
X-Decoder (L)	3×All	0.26M	44.7	13.6±0.3	49.3±0.6	4.2±1.1	8.8±0.4	44.6±0.3	51.1±0.8	72.8±0.6	78.9±1.2	42.0±0.0	13.8±3.1	90.2±0.6	67.8±0.3	11.8±0.1	52.7±0.4	84.3±0.0	18.9±1.2	21.6±1.6	54.3±2.3	56.8±3.5	50.4±0.6	90.7±0.2	42.1±2.7	11.1±1.0	39.7±1.2	44.3±2.9

Table 8. SegInW results with tuning on X-Decoder for different image shots and backbone architectures. (0.26M parameters tuned in the setting.)

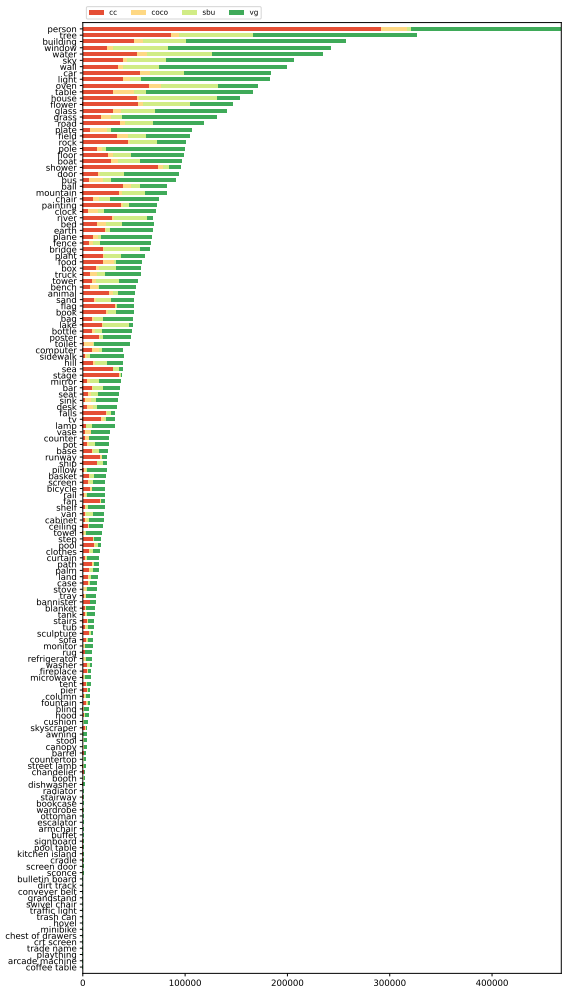


Figure 6. Image captions overlap with ADE20K-150

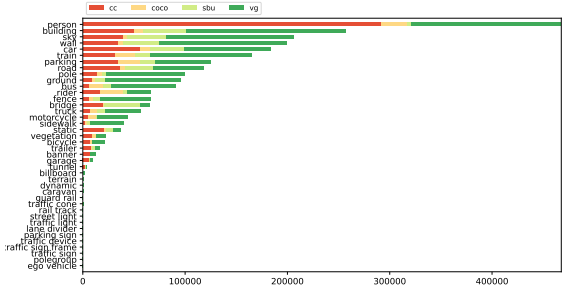


Figure 7. Image captions overlap with BDD-Panoptic

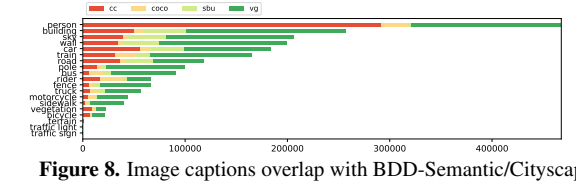


Figure 8. Image captions overlap with BDD-Semantic/Cityscapes

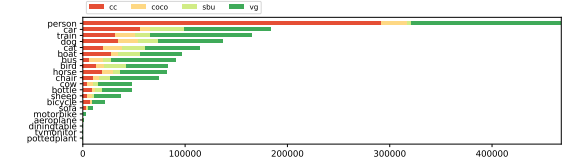


Figure 9. Image captions overlap with Pascal VOC

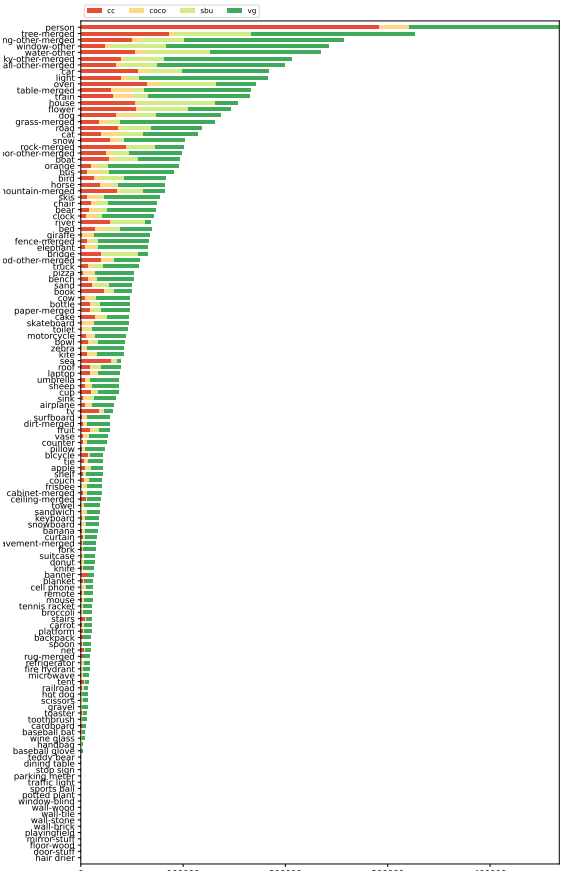


Figure 10. Image captions overlap with COCO

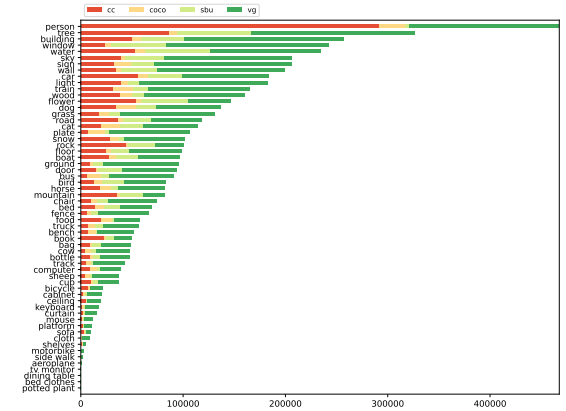


Figure 11. Image captions overlap with Pascal Context-59

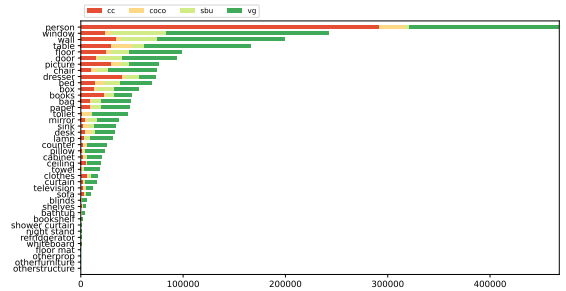


Figure 12. Image captions overlap with ScanNet-40

Figure 13. Zero-Shot Video Generic Segmentation. (Source: YoutubeVOS videos)

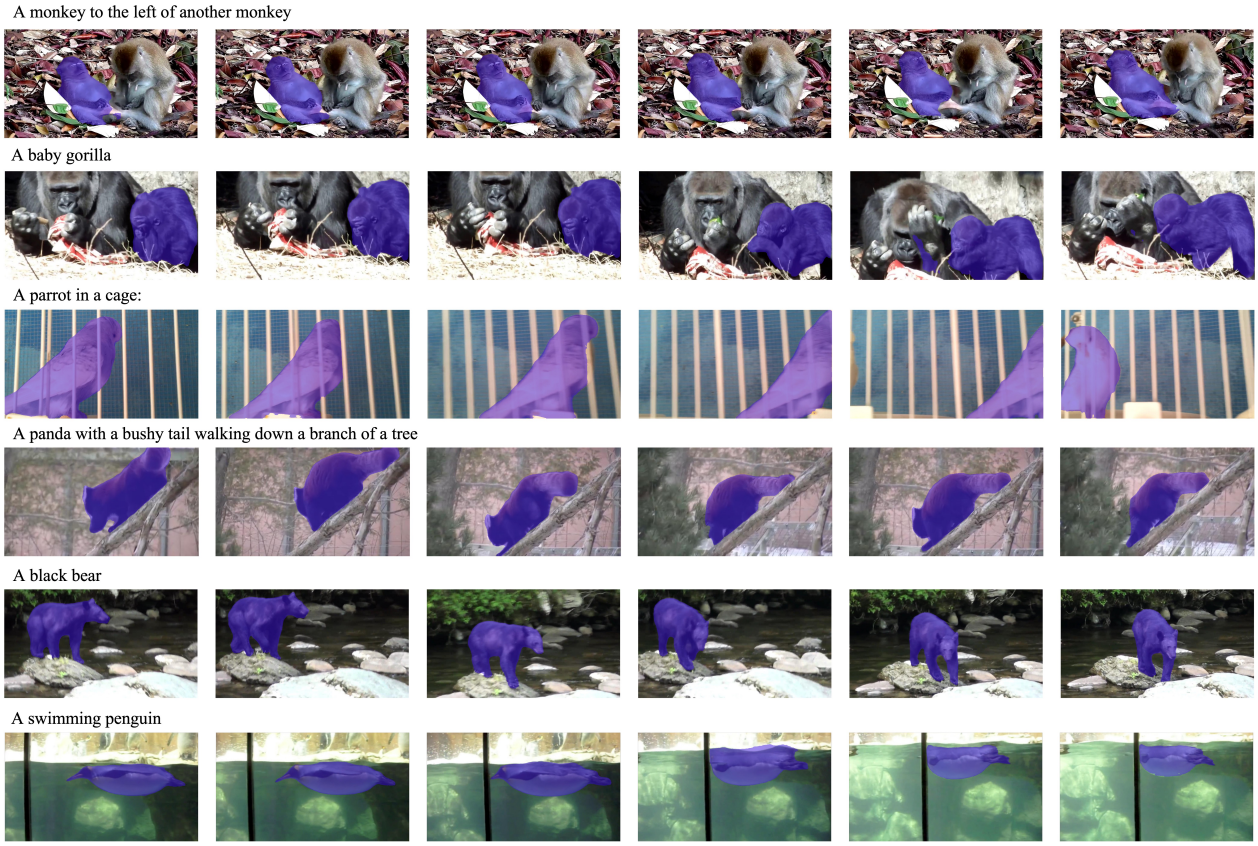


Figure 14. Zero-Shot Referring Video Segmentation. (Source: YoutubeVOS videos)

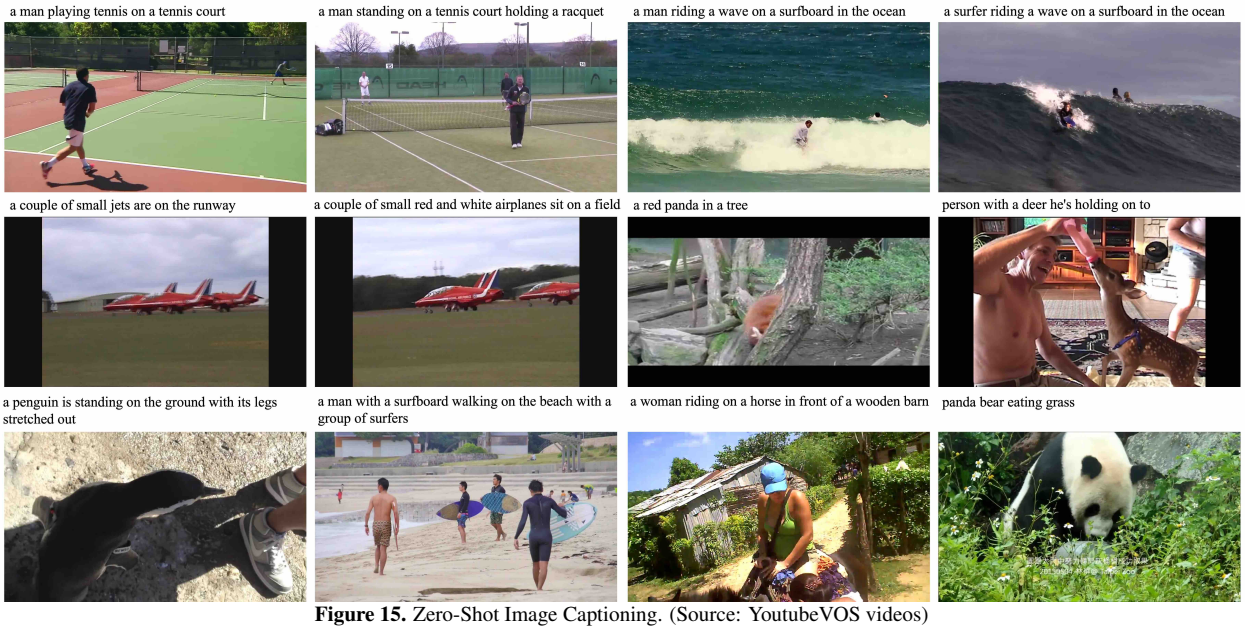


Figure 15. Zero-Shot Image Captioning. (Source: YoutubeVOS videos)

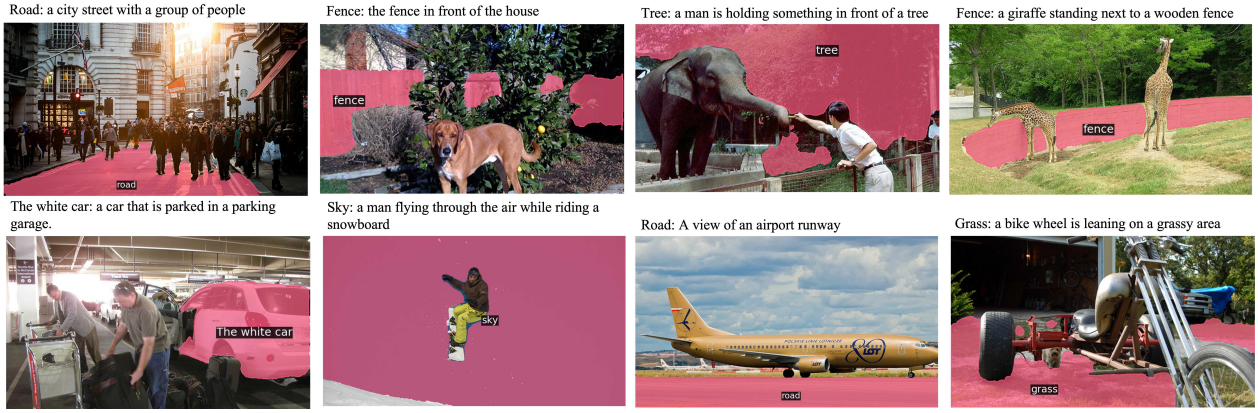


Figure 16. Referring Captioning. (Source: COCO 2017 val images)

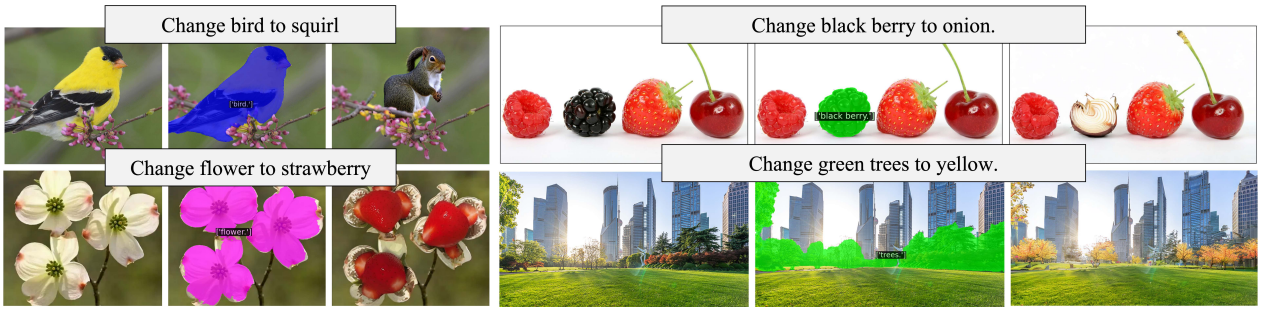


Figure 17. Referring Image Inpainting. (Source: web images)

References

[1] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[3] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. *arXiv preprint arXiv:2204.03645*, 2022.

[4] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[5] 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.

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[9] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.

[10] Jianwei Yang, Chunyuan Li, and Jianfeng Gao. Focal modulation networks. *arXiv preprint arXiv:2203.11926*, 2022.

[11] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *CVPR*, 2022.

[12] Xuwang Yin and Vicente Ordonez. Obj2text: Generating visually descriptive language from object layouts. *arXiv preprint arXiv:1707.07102*, 2017.

[13] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.

[14] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[15] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.