# 1. Supplementary Material

These supplemental materials provide additional details and qualitative results to complement the main manuscript.

## 1.1. Text Prompt For ViP-LLaVa

We retrieve the class name by prompting ViP-LLaVA with the support image and the text prompt:

Prompt 1. ViP-LLaVA Class Name Prompt

```
Human: <image> What is the name of the
   ↪ object inside the red mask
   ↪ contour in the image? Your output
   ↪  must be only the class name of
   ↪ the object. Do not add any
   ↪ additional text.
```

We retrieve the object description of the class name from ViP-LLaVa with the following prompt:

Prompt 2. ViP-LLaVA Class Description Prompt

```
Human: <image> Given the image provided
   ↪ , identify and provide the
   ↪ definition of the {} inside the
   ↪ {} mask contour. Assistant:
```

## 1.2. Visual Prompt For ViP-LLaVa

We prompt ViP-LLaVA on the support image using a mask contour highlight (red, $\alpha = 0.5$, thickness=2) and a 50% zoom. This configuration is compared with competing ones in Table 1) where we compare three prompt types (mask overlay, bounding box, mask contour), varied contour color, transparency (`alpha`), thickness, and zoom level, and measuring a configuration accuracy as:

$$\text{Accuracy} = \frac{\#\{\text{matched names}\}}{\#\{\text{total predictions}\}}.$$

Predicted and ground-truth labels are first lowercased and string-compared; only exact matches count as correct. To have a metric more resilient to synonyms, in the column "Acc. Matching Wordnet Synset(%)" we report the Accuracy scores considering a positive match when the prediction and ground truth share the same set of synsets in WordNet; otherwise, they are incorrect.

**Zoom-based Cropping.** We compute the tightest bounding rectangle around all connected mask components, expand its width and height by the zoom percentage (clipped to image borders), crop, then resize back to the original frame as exemplified in Algorithm 1. When the object already fills most of the image, the result is nearly identical to the original; when the object is small or off-center, the expansion brings it closer into focus so that the vision-language model can attend to it more effectively.

**Algorithm 1** Zoom-based Cropping for ViP-LLaVA Prompting

```
Require: image $I$, mask $M$, zoom $Z\%$
Ensure: zoomed crop $C$
bbox $\leftarrow$ get_bbox($M$)
(bbox.x, bbox.y) $\leftarrow$ (bbox.x$_{\text{top-left}}$, bbox.y$_{\text{top-left}}$)
$(w, h) \leftarrow$ (bbox.width, bbox.height)
bbox.x$_{\text{center}} \leftarrow$ bbox.x$_{\text{top-left}} + w/2$,
bbox.y$_{\text{center}} \leftarrow$ bbox.y$_{\text{top-left}} + h/2$
$w' \leftarrow \min(w \times (100/Z),\ I.\text{width})$
$h' \leftarrow \min(h \times (100/Z),\ I.\text{height})$
bbox.x' $\leftarrow \max(0,\ \text{bbox.x}_{\text{center}} - w'/2)$
bbox.y' $\leftarrow \max(0,\ \text{bbox.y}_{\text{center}} - h'/2)$
$C_{\text{raw}} \leftarrow$ crop($I$, bbox)
$C \leftarrow$ resize($C_{\text{raw}}$, $I$.width, $I$.height)
return $C$
```

## 1.3. Prompt For Visual-Text Alignment Module

The $VTA$ serves as a saliency map that highlights the regions in the query image where the object of interest, described by textual information, is likely located. It leverages the visual-textual alignment capabilities of the CLIP model [1].

The construction of $VTA$ begins with the class name of the entity of interest extracted by ViP-LLaVa. Two text prompts are then generated: a positive prompt $t_{FG}$

Prompt 3. ViP-LLaVA $t_{FG}$ Prompt

```
 a {predicted_class_name}.
```

and a negative prompt $t_{BG}$:

Prompt 4. ViP-LLaVA $t_{BG}$ Prompt

```
 a photo without {predicted_class_name
   ↪ }.
```

## 1.4. MARS Default Configuration

The configuration of the models used in MARS and all the experiments is as follows:

- Text-extraction Module: this component uses a ViP-LLaVA vision-language model based on LLaVA-7B with 4-bit quantization. The prompting strategies employed to extract relevant information from the support set are detailed in subsection 1.1.
- Visual-Text Alignment Module: the proposed method is used to generate $VTA$, computed with a pre-trained CLIP-B/16 model. The threshold parameter in the PIR module, which is applied to refine the initial $VTA$, is set to 0.4 and employs the attention maps extracted from the last 8 self-attention layers (out of the 12) of the encoder, as reported in PI-CLIP [2].

Table 1. Comparison of visual prompt configurations for class name extraction on COCO-$20^i$. "Acc. Exact Matching" reports the percentage of predictions that exactly match the ground-truth label; "Acc. Matching Wordnet Synset" reports the percentage of predictions that match via shared WordNet synsets. The best configuration is highlighted in bold.

| Prompt Type | Color | Alpha | Thickness | Zoom | Acc. Exact Matching (%) | Acc. Matching Wordnet Synset(%) |
|---|---|---|---|---|---|---|
| mask overlay | red | 0.3 | 1 | 0 | 56.37 | 60.74 |
| bounding box | red | 0.5 | 2 | 0 | 62.60 | 68.29 |
| bounding box | red | 0.5 | 2 | 30 | 63.45 | 69.80 |
| bounding box | red | 0.5 | 2 | 50 | 62.82 | 69.07 |
| mask contour | red | 0.5 | 2 | 0 | 62.30 | 68.75 |
| mask contour | red | 0.5 | 2 | 30 | 65.39 | 71.75 |
| mask contour | green | 0.5 | 2 | 50 | 65.70 | 72.23 |
| **mask contour** | **red** | **0.5** | **2** | **50** | **65.75** | **72.25** |

- Visual-Visual Alignment Module: this module employs a Vision Transformer ViT-L/14 pre-trained with DINOv2, specifically the variant using four register tokens, to extract $RVA$. The PIR module threshold used to refine $RVA$ is set to 0.85 and employs the attention maps extracted from all 24 self-attention layers of the encoder.
- AlphaClip Module: a pre-trained AlphaCLIP-L/14@336 model is used to compute the global-conceptual score for each mask proposal.
- Filtering-Merging Module: the fixed threshold thresholdstatic in the filtering component is set to 0.55. The dynamic threshold thresholddynamic is set to 0.95, ensuring that if no proposals exceed the fixed threshold, only those achieving at least 95% of the score of the best proposal are retained.

We have always adopted original weights available from official repositories.

## 1.5. Computational overhead and resource requirements

We assessed the computational overhead introduced by MARS and found it to be minimal when compared to similar pipelines. In particular, we benchmarked the time required for each component of MARS and the competing method Matcher using the COCO-$20^{\text{i}}$ dataset, where hundreds of mask proposals are ranked for a single query. All experiments were run on an NVIDIA Quadro RTX 6000 (24GB), with MARS requiring approximately 18 GB of VRAM to operate.

The initial mask proposal step, performed by SAM, takes approximately $12.1 \pm 1.31$ seconds, which is consistent across both Matcher and MARS. The main distinction lies in the ranking and merging phase. Matcher completes this step in $0.27 \pm 0.25$ seconds, while MARS requires $2.04 \pm 0.61$ seconds. Despite being slower, this additional cost remains negligible relative to the overall computation time dominated by mask generation.

Additionally, in MARS, class name prediction is required only once per support set. This step takes $4.72 \pm 0.59$ seconds and is not repeated for subsequent queries referencing the same support set. Therefore, in the typical scenario where the support set has already been processed, MARS requires approximately 13–14 seconds per prediction, only slightly more than Matcher's 12–13 seconds.

In the worst-case scenario, where the support set is encountered for the first time, the total inference time increases to 18–19 seconds. However, this case occurs only once per novel support set and does not represent the standard operational cost.

## 1.6. Evaluation over novel dataset and classes

A potential limitation of multimodal few-shot segmentation is over-reliance on modules (e.g., ViP-LLaVA) pretrained on generic, web-scale priors and common-class semantics. If true, methods might fail when facing classes underrepresented in vision–language pretraining. To test whether MARS maintains effectiveness in this scenario, we use the refined portion of the ACRE23 dataset [1], a topic-specific dataset for agricultural robotics (released on October 31, 2024; i.e. after the release of ViP-LLaVA) with labeled categories spanning weeds: ryegrass (*Lolium perenne*) and mustard (*Sinapis arvensis*); and crops: corn (*Zea mays*) and beans (*Phaseolus vulgaris*). Discriminative differences of these plants are primarily leaf shape and arrangement, often requiring agronomic knowledge. We evaluate Matcher and MARS+Matcher following the same protocol used in the main paper, plus supplementing WordNet with entries for "corn crop", "bean crop", and "weed grass", detailed in Listing 1. The model demonstrated to correctly use the extended dictionary, improving from 54.17 (Matcher) to 54.44 mIoU (MARS+Matcher).

[1] 10.17632/yxy7drms8k.1

Listing 1. Agronomic additional synsets

```
("corn plants", "small plants with slim
   ↪ , thin, long, curved green leaves
   ↪ "),
("bean plants", "small plants with
   ↪ green heart, fish and pea shaped
   ↪ leaves with clear veins."),
("grass weed", "small scattered plants
   ↪ with thin elongated green blade
   ↪ leaves and pointed tip, pod grass
   ↪ ")
```

## 1.7. SAM Automatic Mask Generation Hyperparameters

In this section, we report the parameters of the *SamAutomaticMaskGenerator* class (taken from the official Matcher repository) used to generate the mask proposals.

```
points_per_side = 64,
points_per_batch = 64,
pred_iou_thresh = 0.88,
sel_pred_iou_thresh = 0.85,
stability_score_thresh = 0.95,
stability_score_offset = 1.0,
sel_stability_score_thresh = 0.90,
sel_stability_score_offset = 1.0,
box_nms_thresh = 0.65,
crop_n_layers = 1,
crop_nms_thresh = 0.7,
crop_overlap_ratio = 512 / 1500,
crop_n_points_downscale_factor = 2,
point_grids = None,
min_mask_region_area = 14,
output_mode = "binary_mask",
multimask_output = True,
sel_multimask_output = True,
output_layer = -1,
sel_output_layer = -1,
dense_pred = True
```

## 1.8. Baseline implementation details

We have reevaluated the baseline presented in Tables **??** and **??** with the following details:

### PerSAM

- **Weights** The model is based on SAM. The used version is reported below.
- **Backbone** SAM ViT-H.

### VRP-SAM

- **Weights** Official checkpoint not released. We trained the model using the provided training script.
- **Backbone** ResNet50.

### SegGPT

- **Weights** Official released checkpoint.

### Matcher

- **Weights** The model is based on different backbones; the used version is reported below.
- **Backbone** DINOv2 ViT-L/14, SAM ViT-H

### GF-SAM

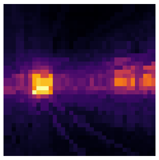
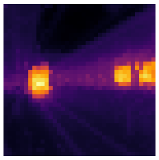
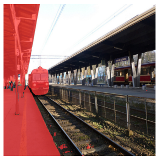
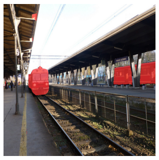
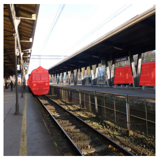
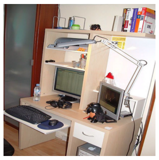
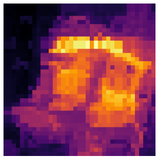
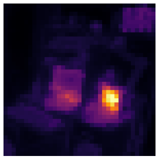
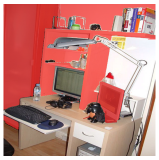
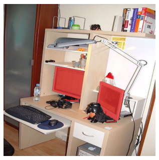
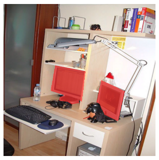
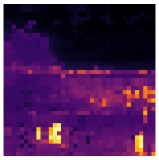
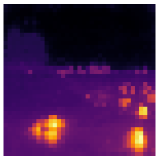
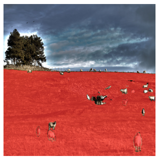
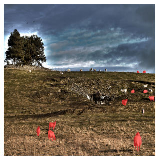
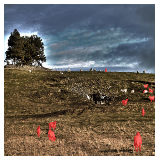
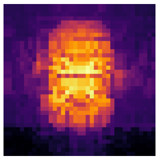
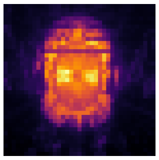
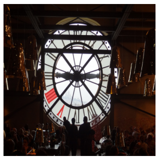
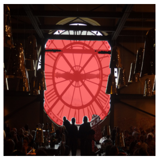
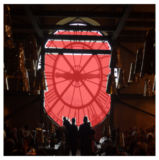
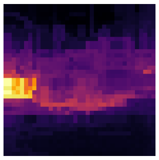
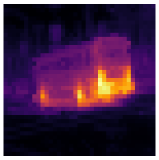
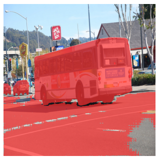
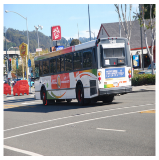
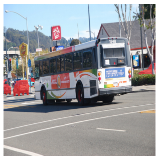
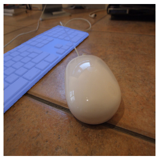
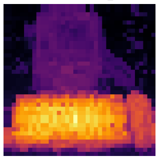
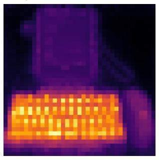
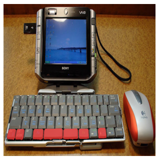
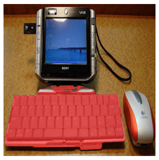
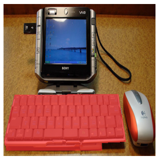
- **Weights** The model is based on different backbones; the used version is reported below.
- **Backbone** DINOv2 ViT-L/14, SAM ViT-H

### FSSAM

- **Weights** Official released checkpoint.
- **Backbone** Small based on SAM2 small.
- **Notes** We use the provided list of episodes for evaluation.

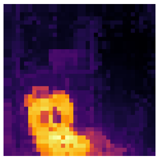
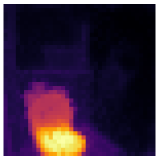
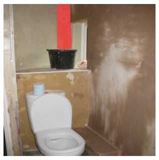
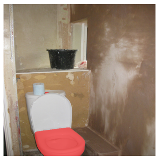
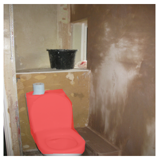
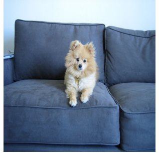
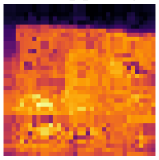
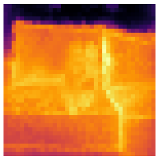
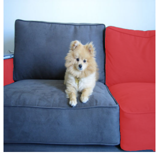
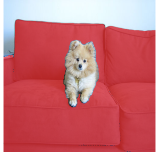
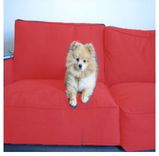
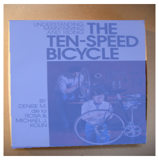
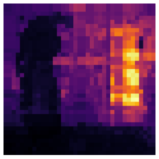
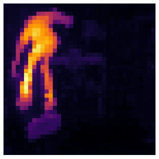
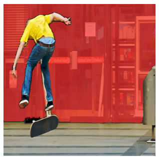
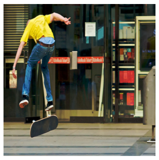
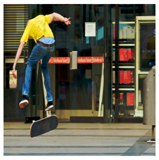
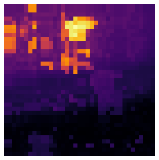
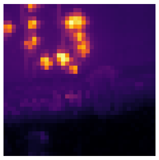
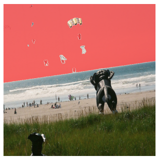
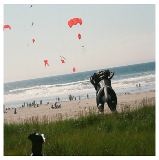
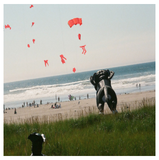
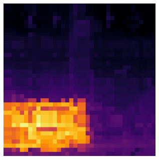
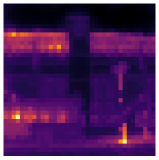
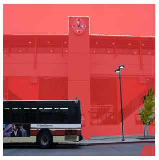
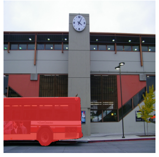
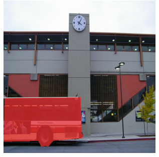
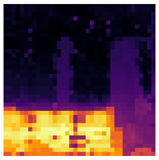
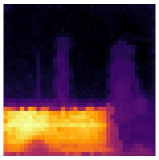
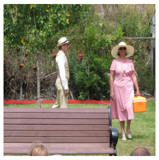
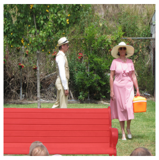
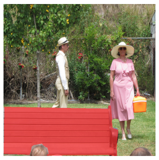
## 1.9. Qualitative Results

Table 2. Qualitative Results: This table presents each sample as a two-part row. The upper part displays the Support Set, Query Image, Visual Prior, Text Prior, Matcher Prediction, Matcher with MARS Prediction, and the Ground Truth. The lower part provides the corresponding textual information: the dataset's class name, the predicted class name, a description of the subject class inferred by MARS using WordNet, followed by the IoU values for the Matcher prediction and the Matcher with MARS prediction.

| Support Set | Query Image | Text Prior | Visual Prior | Matcher | Matcher + MARS | Ground Truth |
|---|---|---|---|---|---|---|
| | |  |  |  |  |  |
| train | Train | *Description not found* | | 7.38 mIoU | 87.11 mIoU | |
| |  |  |  |  |  |  |
| tv | Television | *an electronic device that receives television signals and displays them on a screen* | | 7.59 mIoU | 86.63 mIoU | |
| | |  |  |  |  |  |
| sheep | Sheep | *woolly usually horned ruminant mammal related to the goat* | | 1.67 mIoU | 64.91 mIoU | |
| | |  |  |  |  |  |
| clock | Clock | *a timepiece that shows the time of day* | | 2.94 mIoU | 85.34 mIoU | |
| | |  |  |  |  |  |
| car | Car | *Description not found* | | 1.96 mIoU | 65.84 mIoU | |
|  | |  |  |  |  |  |
| keyboard | Keyboard | *Description not found* | | 12.76 mIoU | 89.25 mIoU | |

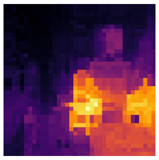
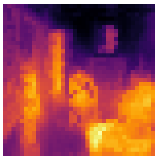
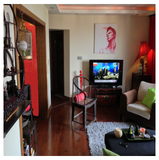
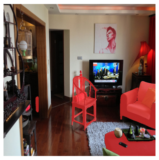
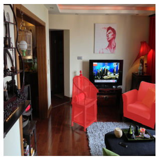
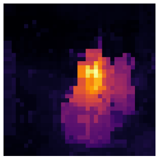
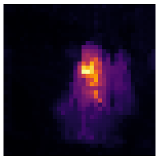
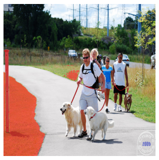
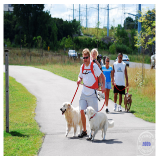
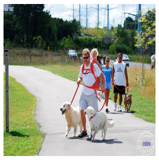
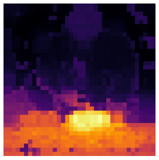
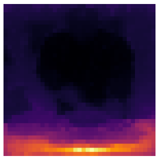
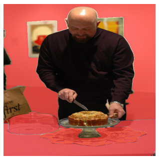
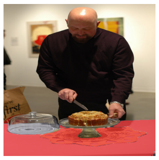
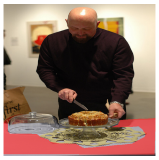
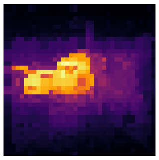
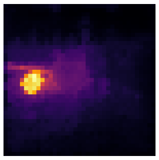
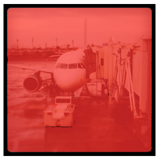
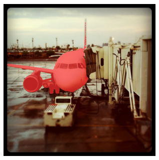
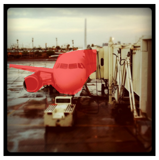
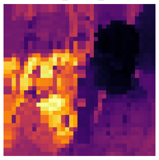
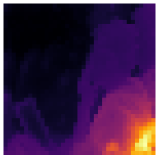
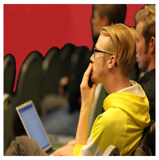
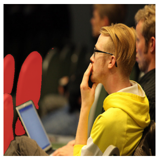
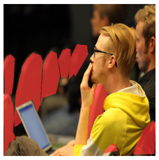
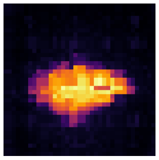
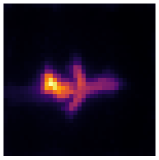
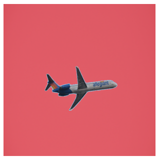
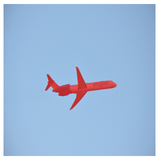
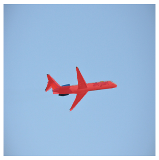
Continued on next page

Table 2 – continued from previous page

| Support Set | Query Image | Text Prior | Visual Prior | Matcher | MARS | Ground Truth |
|---|---|---|---|---|---|---|
| | |  |  |  |  |  |
| toilet | Chair | *a seat for one person, with a support for the back* | | 0.13 mIoU | 38.24 mIoU | |
| |  |  |  |  |  |  |
| couch | Couch | **Description not found** | | 33.31 mIoU | 92.98 mIoU | |
|  | |  |  |  |  |  |
| book | Book | *a written work or composition that has been published (printed on pages bound together)* | | 2.47 mIoU | 51.30 mIoU | |
| | |  |  |  |  |  |
| kite | Kite | *any of several small graceful hawks of the family Accipitridae having long pointed wings and feed...* | | 0.74 mIoU | 59.23 mIoU | |
| | |  |  |  |  |  |
| bus | Bus | *a vehicle carrying many passengers; used for public transport* | | 0.02 mIoU | 95.65 mIoU | |
| | |  |  |  |  |  |
| bench | Bench | *the magistrate or judge or judges sitting in court in judicial capacity to compose the court coll...* | | 0.00 mIoU | 94.98 mIoU | |

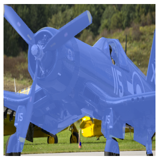
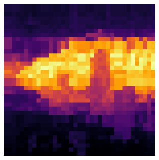
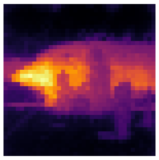
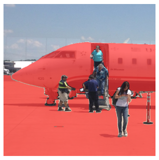
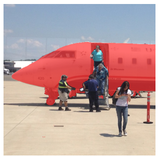
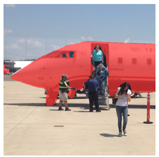
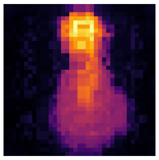
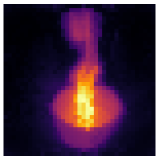
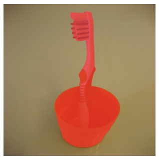
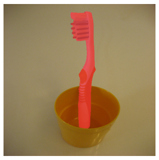
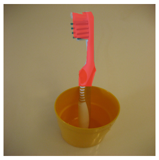
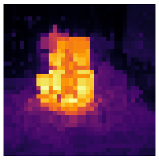
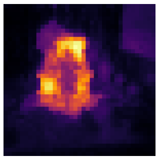
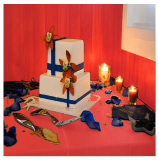
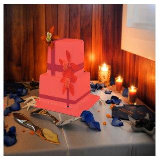
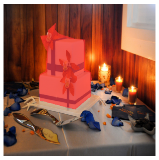
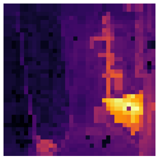
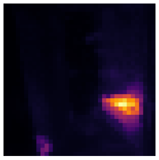
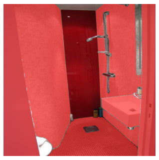
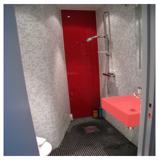
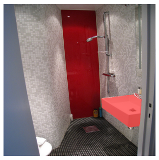
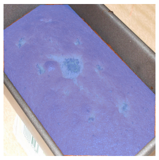
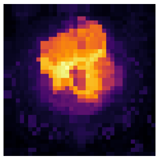
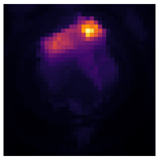
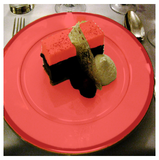
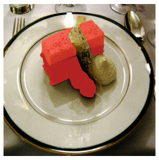
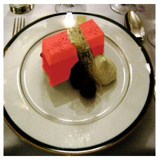
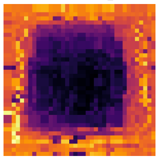
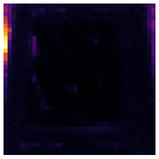
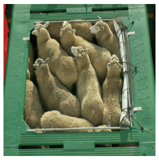
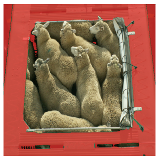
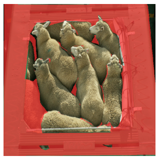
Continued on next page

Table 2 – continued from previous page

| Support Set | Query Image | Text Prior | Visual Prior | Matcher | MARS | Ground Truth |
|---|---|---|---|---|---|---|
| | |  |  |  |  |  |
| chair | Seat | *any support where you can sit (especially the part of a chair or bench etc. on which you sit)* | | 0.00 mIoU | 57.79 mIoU | |
| | |  |  |  |  |  |
| backpack | Backpack | *a bag carried by a strap on your back or shoulder* | | 0.00 mIoU | 58.18 mIoU | |
| | |  |  |  |  |  |
| dining table | Table | *a company of people assembled at a table for a meal or game* | | 19.90 mIoU | 58.23 mIoU | |
| | |  |  |  |  |  |
| airplane | Airplane | *an aircraft that has a fixed wing and is powered by propellers or jets* | | 10.36 mIoU | 72.47 mIoU | |
| | |  |  |  |  |  |
| chair | Chair | **Description not found** | | 0.17 mIoU | 40.51 mIoU | |
| | |  |  |  |  |  |
| airplane | Jet engine | *a gas turbine produces a stream of hot gas that propels a jet plane by reaction propulsion* | | 0.12 mIoU | 83.25 mIoU | |

Continued on next page

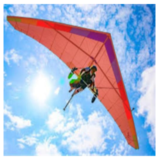
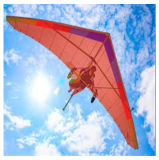
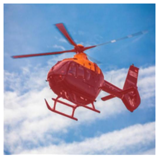
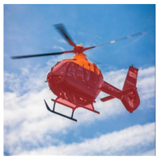
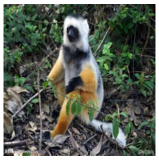
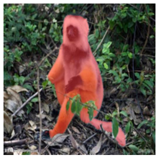
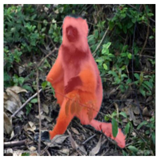
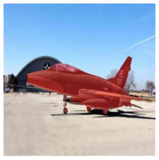
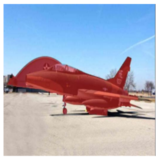
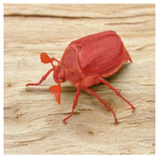
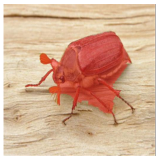
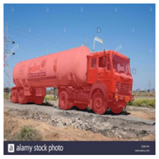
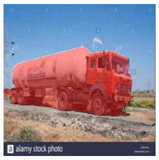
Table 2 – continued from previous page

| Support Set | Query Image | Text Prior | Visual Prior | Matcher | MARS | Ground Truth |
|---|---|---|---|---|---|---|
| <br>airplane | <br>Airplane | <br>*an aircraft that has a fixed wing and is powered by propellers or jets* |  | <br>35.28 mIoU | <br>91.19 mIoU |  |
| <br>toothbrush | <br>Toothbrush | <br>*small brush; has long handle; used to clean teeth* |  | <br>23.30 mIoU | <br>59.88 mIoU |  |
| <br>cake | <br>Cake | <br>*baked goods made from or based on a mixture of flour, sugar, eggs, and fat* |  | <br>0.87 mIoU | <br>74.98 mIoU |  |
| <br>sink | <br>Sink | <br>**Description not found** |  | <br>5.07 mIoU | <br>94.34 mIoU |  |
| <br>cake | <br>Cake | <br>**Description not found** |  | <br>8.21 mIoU | <br>73.46 mIoU |  |
| <br>truck | <br>Truck | <br>*an automotive vehicle suitable for hauling* |  | <br>1.51 mIoU | <br>75.79 mIoU |  |

## 1.10. Qualitative inspection of FSS-1000

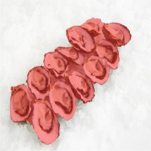
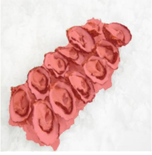
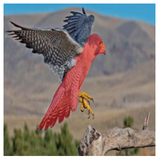
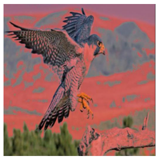
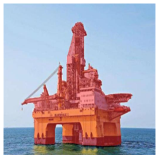
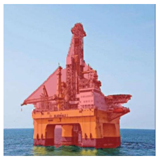
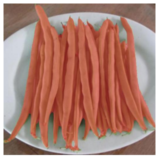
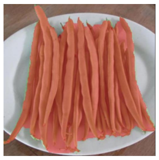
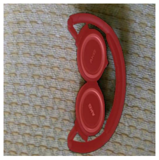
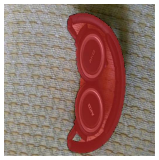
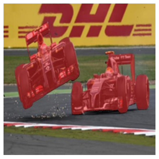
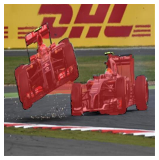
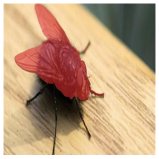
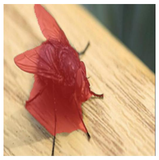
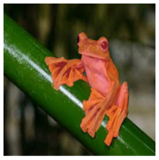
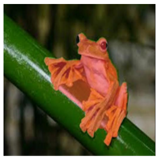
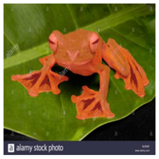
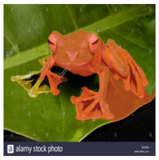
The results of our experiments in Tables **??** and **??** show small or even negative improvements on FS-1000, in contrast to the more consistent gains observed on the other datasets. For this reason, we investigate these experiments in greater depth. Several cases reveal noisy or imperfect ground truth in FSS-1000: fine details sometimes marked as subject are actually background on closer inspection. This label noise penalizes otherwise improved masks, so surpassing already strong models on this dataset can be difficult since better predictions may slightly degrade IoU when compared against imperfect labels. Table 3 shows a selection of such cases where Matcher + MARS predictions are qualitatively better than the ground truth but achieve suboptimal IoU scores.

Table 3. Label quality analysis on FSS-1000 with MARS + Matcher prediction. Each row displays, left to right, the Support image, the Query image, the model Prediction with its IoU, and the Ground Truth. Several examples illustrate that subtle background structures are labeled as foreground, which can reduce IoU for more precise predictions.

| Support Set | Query Image | Prediction | Ground Truth |
|---|---|---|---|
| | | <br>87.70 IoU |  |
| | | <br>85.57 IoU |  |
| |  | <br>86.23 IoU |  |
| | | <br>76.41 IoU |  |
| | | <br>72.65 IoU |  |
| | | <br>92.88 IoU |  |

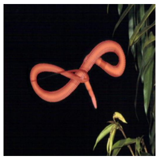
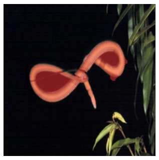
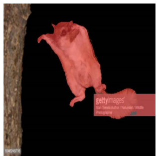
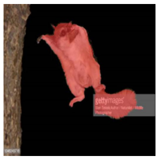
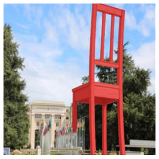
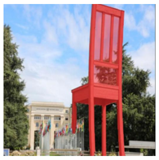
Continued on next page

Table 3 continued

| Support Set | Query Image | Prediction | Ground Truth |
|---|---|---|---|
| | | <br>88.59 IoU |  |
| | | <br>5.75 IoU |  |
| | | <br>88.56 IoU |  |
| | | <br>89.34 IoU |  |
| | | <br>81.39 IoU |  |
| | | <br>87.87 IoU |  |
| | | <br>59.85 IoU |  |
| | | <br>83.63 IoU |  |
| | | <br>70.23 IoU |  |

Continued on next page

Table 3 continued

| Support Set | Query Image | Prediction | Ground Truth |
|---|---|---|---|
| | | <br>67.17 IoU |  |
| | | <br>83.68 IoU |  |
| | | <br>70.90 IoU |  |

## References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 1

[2] Jin Wang, Bingfeng Zhang, Jian Pang, Honglong Chen, and Weifeng Liu. Rethinking prior information generation with CLIP for few-shot segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 3941–3951. IEEE, 2024. 1