Appendix

In this supplementary material, we provide more details about the datasets we use, implementation details and ablations, as well as further qualitative and quantitative evaluations.

1. Datasets

As noted in the main paper, we contribute additional annotations to the Spair71k dataset for some of our experiments. We start from their keypoint annotations, which have no keypoint name annotations in the original dataset. We then manually name all keypoints of the animal classes in Spair71k, as shown in Table 2. We purposefully leave out some point annotations:

- All animals have a left and right nostril annotated — we take the right one in all classes and annotate it as *nose*, and leave the left nostril out.
- All tails have point annotations at the start of the tail (attached to the body) and end of the tail. Because of the lack of words to precisely describe both points, we take the point *not* attached to the body and annotate it as *tail*, and leave the other one out.
- All ears have point annotations at the start of the ear (attached to the head), and at the pointy end. Because of the lack of words to precisely describe both points, we take the point *not* attached to the head and annotate it as *ear*, and leave the other one out.
- Birds have annotations for (i) foot, (ii) ankle, (iii) knee, which are often ambiguous and very close together. We only keep the *foot* annotation.

Note we explicitly define different names for keypoints that can be ambiguous, e.g. eyes, ears, legs, etc. This ensures the role of questions and answers in **??** is satisfied.

2. Discovered annotations

Out of the discovered annotations in YFCC-15M, 44% contain red circles. Overall, 73% of the annotations were circles, and the rest were rectangles. 65% of all annotations were red, 10% yellow, 7% blue, 7% white, and the rest were black, green, and purple.

3. Additional implementation details

3.1. Referring Expressions Detection.

Backbone We base the evaluation of our method on ReCLIP [1], where an ensemble of two CLIP backbones is used — RN50x16 and ViT-B/32. We evaluated ReCLIP for all combinations of CLIP backbones in Table 1 and found that, on average, this is the highest-performing one. Similarly, for our method, we choose the ensemble of two backbones that lead to the highest performance — RN50x16 and ViT-L/14@336. Full comparison between the backbones can be found in Table 1.

Annotations We experiment with different marker shapes, sizes, and colours, and present the results in Table 4. We find that, on average, a thin red circle leads to the best performance. We use an ensemble of the red circle annotation and two additional augmentations — blurring and gray-scaling the outside of the circle, for a total of three images per annotation, as shown. These augmentations were inspired by examples in YFCC15M we discovered that were annotated like that. We found that adding augmentations improves overall results. However, we did not explore including augmentations beyond these. We ablate these choices in Table 3.

Additional details We augment the text queries by prepending *"This is"*. When subtracting the average with respect to other referring expressions, we use $Q = 500$ randomly sampled expressions.

3.2. Keypoint tasks

Backbone We evaluate different backbones in Table 3 in the main paper and find that ViT-L/14@336 performs best.

Annotations We show examples of the markers we use in Fig. 4 in the main paper . We compare a large range of sizes and colors, as shown in Table 2 in the main paper. We find that a circle is the best marker, and drawing a cross over the point of interest is the worst. The best-performing marker out of all is a red circle, which is the one we end up using. In Fig. 2 we show a more detailed comparison of different colors, diameters, and thicknesses when using a circle annotation. We see that a thin red circle is the best-performing marker. We show what that circle looks like on an image in Fig. 3.

Given this, we draw red circles over the images, with radius $r = 0.06H$ and thickness $t = 0.01H$, where H is the shorter side of the image. For the backbone we use, where the input size has $H = 336$px, this becomes $r = 20$px and $t = 3$px.

Additional details For the keypoint localization task, we set $M = 30$, for a total of $30 \times 30 = 900$ query locations before applying the pseudo mask. The templates we use are *"This is the* $\{part\}$ *of a bird"* for CUB and *"This image shows the* $\{part\}$ *of the* $\{animal\}$*"* for SPair71k. We use a temperature parameter $\tau = \frac{1}{150}$.

4. Qualitative evaluations

We present qualitative evaluations on naming keypoints in Figs. 6 and 7, keypoint localization in Figs. 3 and 4 and referring expressions comprehension in Fig. 5.

Method	Backbone	RefCOCO			RefCOCO+			RefCOCOg	
		Val	TestA	TestB	Val	TestA	TestB	Val	Test
ReCLIP	RN50×16	37.61	38.32	37.19	44.12	46.02	41.81	55.94	54.36
	ViT-B/32	40.69	43.98	37.55	45.00	48.15	41.65	55.25	54.35
	ViT-B/16	38.23	40.53	37.00	41.53	42.91	41.32	55.19	55.16
	ViT-L/14	34.40	33.52	34.35	37.86	37.53	37.70	53.82	52.25
	ViT-L/14@336px	35.90	37.72	35.66	40.06	42.49	39.07	54.25	53.92
	RN50×16,ViT-B/32	**41.96**	43.52	39.00	**47.44**	**50.11**	**43.93**	57.76	**57.15**
	RN50×16,ViT-B/1	39.94	41.61	38.71	45.06	47.17	43.63	57.93	56.85
	RN50×16,ViT-L/14	37.98	38.08	37.51	42.87	44.57	41.66	56.78	56.02
	RN50×16,ViT-L/14@336px	38.79	39.49	37.82	44.27	46.44	42.46	57.86	56.28
	ViT-B/32,ViT-B/16	41.34	**44.25**	38.55	45.20	48.01	43.36	57.37	56.52
	ViT-B/32,ViT-L/14	39.68	41.65	37.84	43.74	46.25	41.17	56.74	56.07
	ViT-B/32,ViT-L/14@336px	40.82	43.47	**39.22**	45.41	48.52	42.83	**58.09**	56.94
	ViT-B/16,ViT-L/14	37.69	38.29	37.53	40.87	42.07	40.93	56.35	55.76
	ViT-B/16,ViT-L/14@336px	39.18	41.01	38.35	42.81	44.32	42.07	57.82	56.21
	ViT-L/14,ViT-L/14@336px	35.47	36.26	35.70	39.52	40.69	38.70	54.51	54.04
Red Circle	RN50×16	45.52	52.99	38.59	49.98	57.55	42.11	53.94	54.35
	ViT-B/32	38.72	45.09	33.52	42.85	49.46	36.53	45.81	45.57
	ViT-B/16	45.30	52.70	36.51	49.39	57.67	40.60	53.72	53.26
	ViT-L/14	46.71	55.03	39.24	52.07	58.63	42.83	57.00	56.40
	ViT-L/14@336	48.27	56.44	39.71	53.59	59.99	43.28	**59.95**	58.51
	RN50×16, ViT-B/32	45.62	54.04	37.13	50.73	60.46	41.69	54.00	53.84
	RN50×16, ViT-B/16	**49.98**	57.15	38.04	52.98	61.95	42.99	56.01	55.78
	RN50×16, ViT-L/14	48.50	58.03	39.76	54.56	63.17	44.41	58.17	57.76
	RN50×16, ViT-L/14@336	49.84	**58.57**	39.96	55.28	**63.92**	**45.35**	59.40	**58.93**
	ViT-B/32,ViT-B/16	44.62	53.03	35.90	49.13	58.96	40.21	52.23	51.61
	ViT-B/32,ViT-L/14	47.19	56.27	38.14	52.75	62.07	42.69	56.66	55.54
	ViT-B/32,ViT-L/14@336px	48.59	58.05	38.69	54.61	63.45	43.28	57.80	57.48
	ViT-B/16,ViT-L/14	48.18	57.49	39.33	53.66	62.38	43.36	57.56	57.45
	ViT-B/16,ViT-L/14@336px	49.86	58.41	39.92	**55.35**	62.43	44.34	59.05	58.82
	ViT-L/14,ViT-L/14@336px	48.82	57.03	**40.35**	53.62	60.65	44.04	59.03	58.27

Table 1: Backbone ablation on **Referring Expressions Detection.** We compare CLIP backbones and their ensembles for ReCLIP [1] (without using relations resolution) and our Red Circle. The best and second best for each method are **bolded** and underlined, respectively.

Figure 1: **Annotations for Referring Expressions Detection.** Here we show the annotation types we consider. A: original bounding box annotation. B: Red Circle. C: Red Circle + Blur outside. D: Red Circle + Gray outside. In our experiments, we use an ensemble of B, C and D unless stated otherwise.

Part No	Bird	Cat	Cow	Dog	Horse	Sheep
0	crown	—	—	—	—	—
1	right wing	—	—	—	—	—
2	left wing	right ear	right ear	right ear	right ear	right ear
3	beak	left ear	left ear	left ear	left ear	left ear
4	—	right eye	right eye	right eye	right eye	right eye
5	—	left eye	left eye	left eye	left eye	left eye
6	forehead	nose	nose	nose	nose	nose
7	right eye	—	—	forehead	—	—
8	left eye	mouth	mouth	mouth	mouth	mouth
9	nape	front right paw	front right hoof	front right paw	forehead	front right hoof
10	right foot	front left paw	front left hoof	front left paw	front right hoof	front left hoof
11	left foot	hind right paw	hind right hoof	hind right paw	front left hoof	hind left hoof
12	—	hind left paw	hind left hoof	hind left paw	hind right hoof	hind right hoof
13	tail	tail	tail	tail	hind left hoof	tail
14	—	—	—	—	tail	—
15	—	—	front right knee	neck	—	front right knee
16	—	—	front left knee	—	front right knee	front left knee
17	—	—	hind right knee	—	front left knee	hind right knee
18	—	—	hind left knee	—	hind right knee	hind left knee
19	—	—	right horn	—	hind left knee	right horn
20	—	—	left horn	—	—	—

Table 2: **Part names for keypoint annotations of the SPair71k dataset**. Part No is the part number in the SPair71k annotations. Some parts are annotated inconsistently in the original annotations, e.g. "tail" is part number 10 for the "horse" class, but part number 9 for all other animal classes.

Component			RefCOCO			RefCOCO+			RefCOCOg	
Red Circle	Subtract	Ensemble	Val	TestA	TestB	Val	TestA	TestB	Val	Test
✓	✗	✗	42.01	48.58	36.90	47.55	53.56	41.05	50.84	51.47
✓	✓	✗	43.67	50.20	38.59	48.98	54.70	43.06	54.29	52.98
✓	✓	✓	**49.84**	**58.57**	**39.96**	**55.28**	**63.92**	**45.35**	**59.40**	**58.93**

Table 3: **Ablation study.** We ablate subtracting the mean wrt negative queries and ensembling different marker types (red circle + red circle and blur outside + red circle and grey outside). Here we use RN50×16 and ViT-L/14@336px backbones and a red circle with the optimal size described in Table 4

References

[1] Sanjay Subramanian, Will Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. *arXiv preprint arXiv:2204.05991*, 2022. 1, 2

Annotation Type			RefCOCO			RefCOCO+			RefCOCOg	
Shape	Color	Size	Val	TestA	TestB	Val	TestA	TestB	Val	Test
Circle	Red	1	38.7	<u>45.1</u>	<u>34.0</u>	<u>44.4</u>	**50.0**	<u>39.1</u>	48.1	**50.0**
Circle	Red	2	32.2	35.9	29.1	37.6	40.9	33.5	45.3	46.4
Circle	Red	4	37.4	43.6	31.5	43.3	47.8	37.3	43.7	48.0
Circle	Red	8	36.3	42.6	31.3	42.1	47.3	36.3	45.2	45.4
Rectangle	Red	1	35.1	38.3	33.5	39.2	41.4	37.3	44.3	43.4
Rectangle	Red	2	35.1	38.3	33.2	39.1	41.8	37.3	44.8	44.1
Rectangle	Red	4	34.1	37.8	32.3	39.0	41.3	36.5	43.7	44.1
Rectangle	Red	8	33.7	37.6	32.7	37.9	40.3	34.9	41.1	40.1
Circle	Green	1	**39.3**	**45.4**	**34.8**	43.8	<u>49.9</u>	38.1	47.2	47.4
Circle	Purple	1	<u>38.9</u>	44.8	34.0	**44.5**	49.4	**39.2**	**49.5**	<u>49.2</u>
Circle	Blue	1	37.7	44.9	33.5	43.4	49.1	37.3	48.2	48.3
Circle	Yellow	1	38.5	44.1	34.6	43.7	49.0	38.9	<u>48.6</u>	48.1

Table 4: **Comparison of different sizes, shapes, colors.** A unit size of 1 corresponds to 0.5% of the larger side of the image, which is 1 pixel for an image of size 224. Here we do *not* use ensembling and subtraction of the mean wrt other queries in order to evaluate the effectiveness of different markers themselves. The best and second best are **bolded** and <u>underlined</u>, respectively.

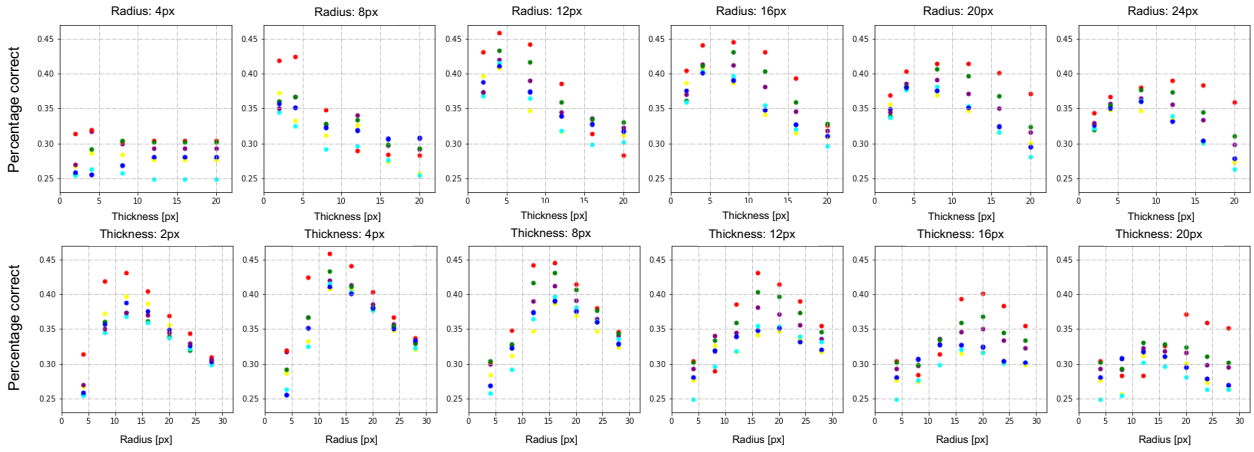

Figure 2: **Ablation of circle sizes and colours for keypoint matching.** We present results on the CUB dataset when varying the diameter and thickness of the annotations. The presented numbers are for text-to-image matching. The best performing annotation has a radius of 12px and thickness of 4px. The colour of the dots on the scatter plot illustrates the colour of the annotation — red, green, blue purple, yellow, cyan.

Figure 3: **Qualitative evaluation of keypoint localization on SPair71k.** We show all keypoint names for the images and color code in green and red (dashed) the correct and wrong localizations according to PCK with $\alpha = 0.1$. The red circle is the marker we use and the blue dot is the ground truth location.

Figure 4: **Qualitative evaluation of keypoint localization on SPair71k.** We show all keypoint names for the images and color code in green and red (dashed) the correct and wrong localizations according to PCK with $\alpha = 0.1$. The red circle is the marker we use and the blue dot is the ground truth location.

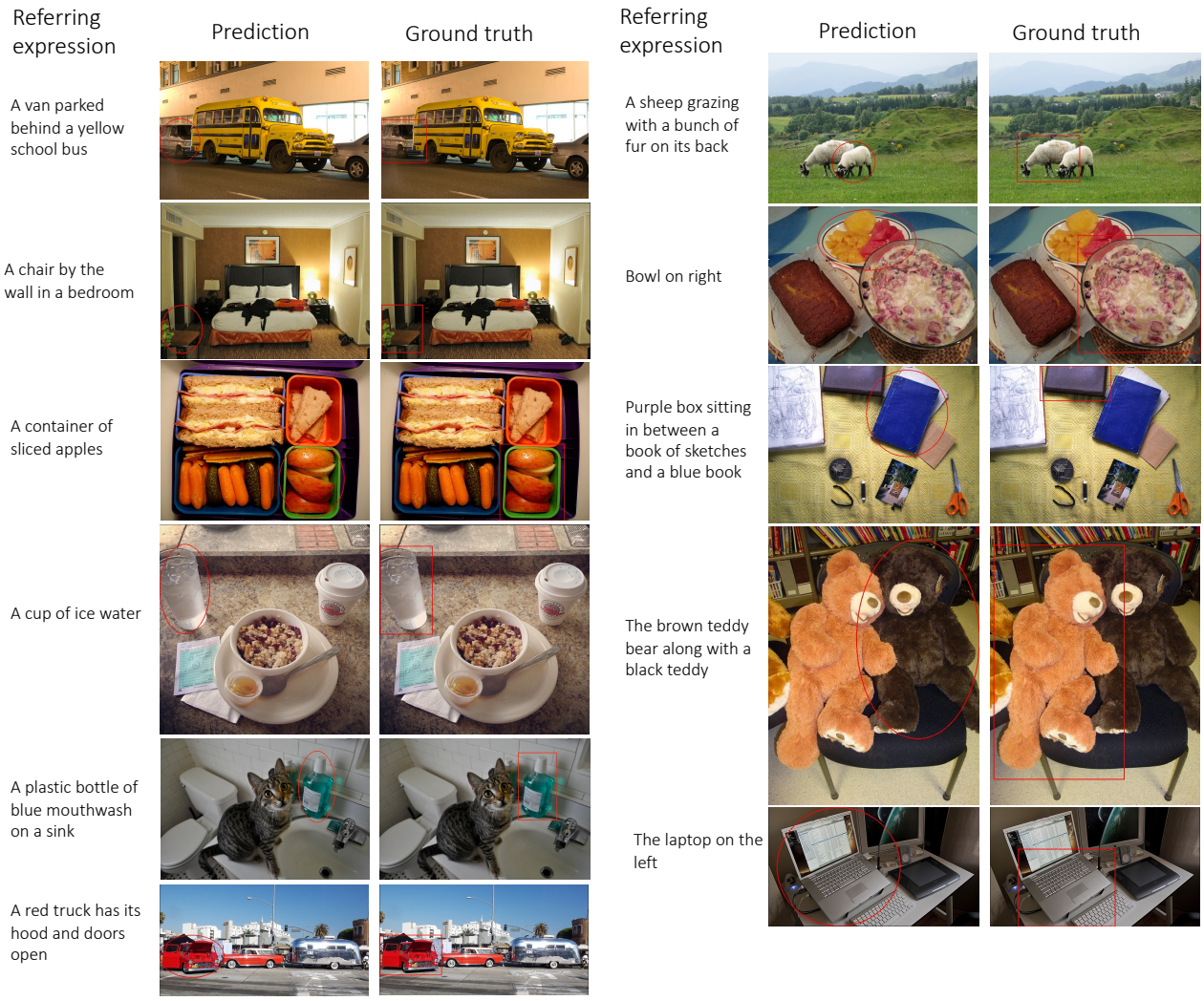

Figure 5: **Qualitative results on REC on the RefCOCOg dataset.** Left: correct predictions. Right: wrong predictions. The last row on the right shows an example where the ground-truth bounding box is wrong.

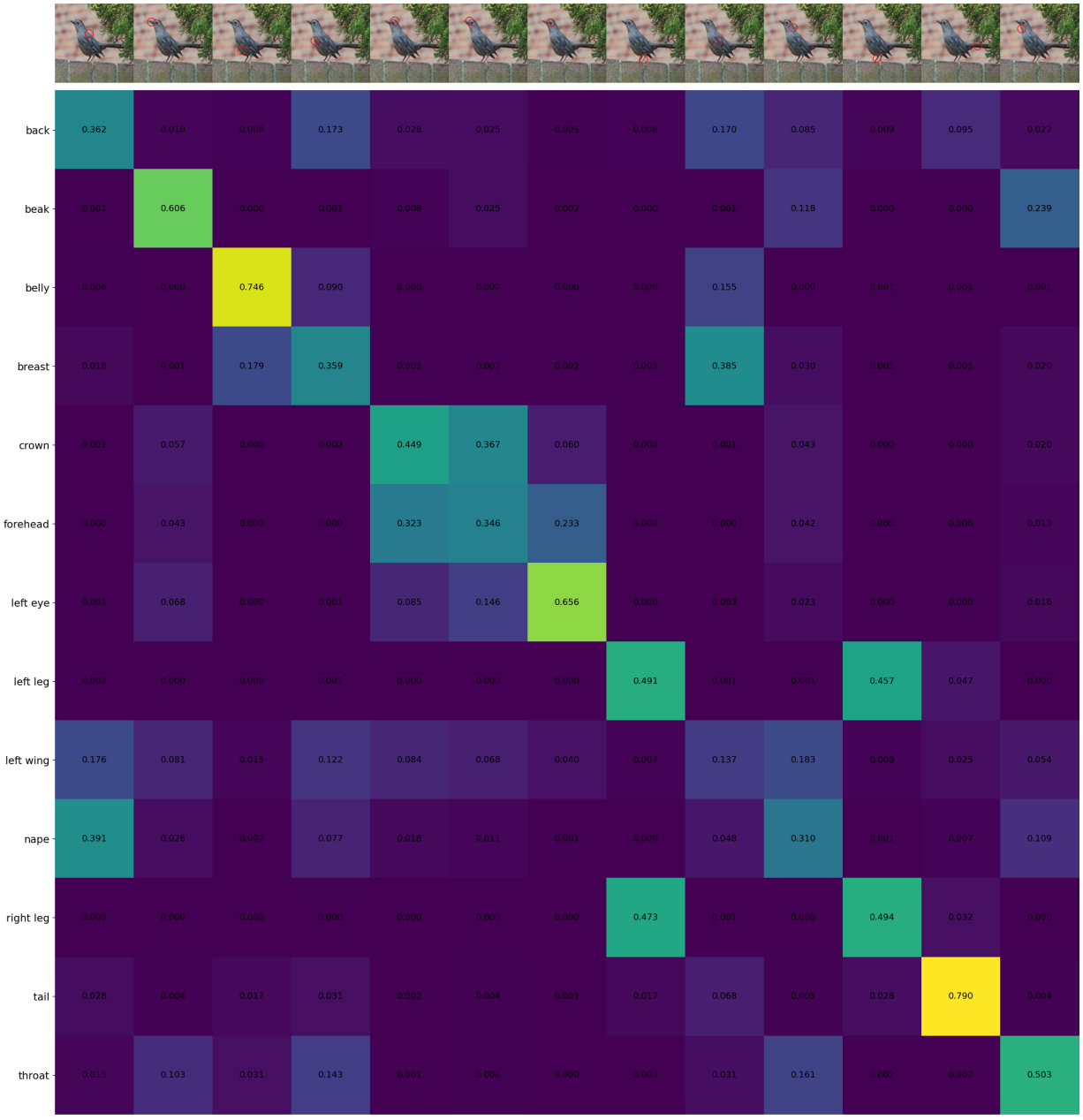

Figure 6: **Naming keypoints.**Normalized cost matrix for an image from CUB

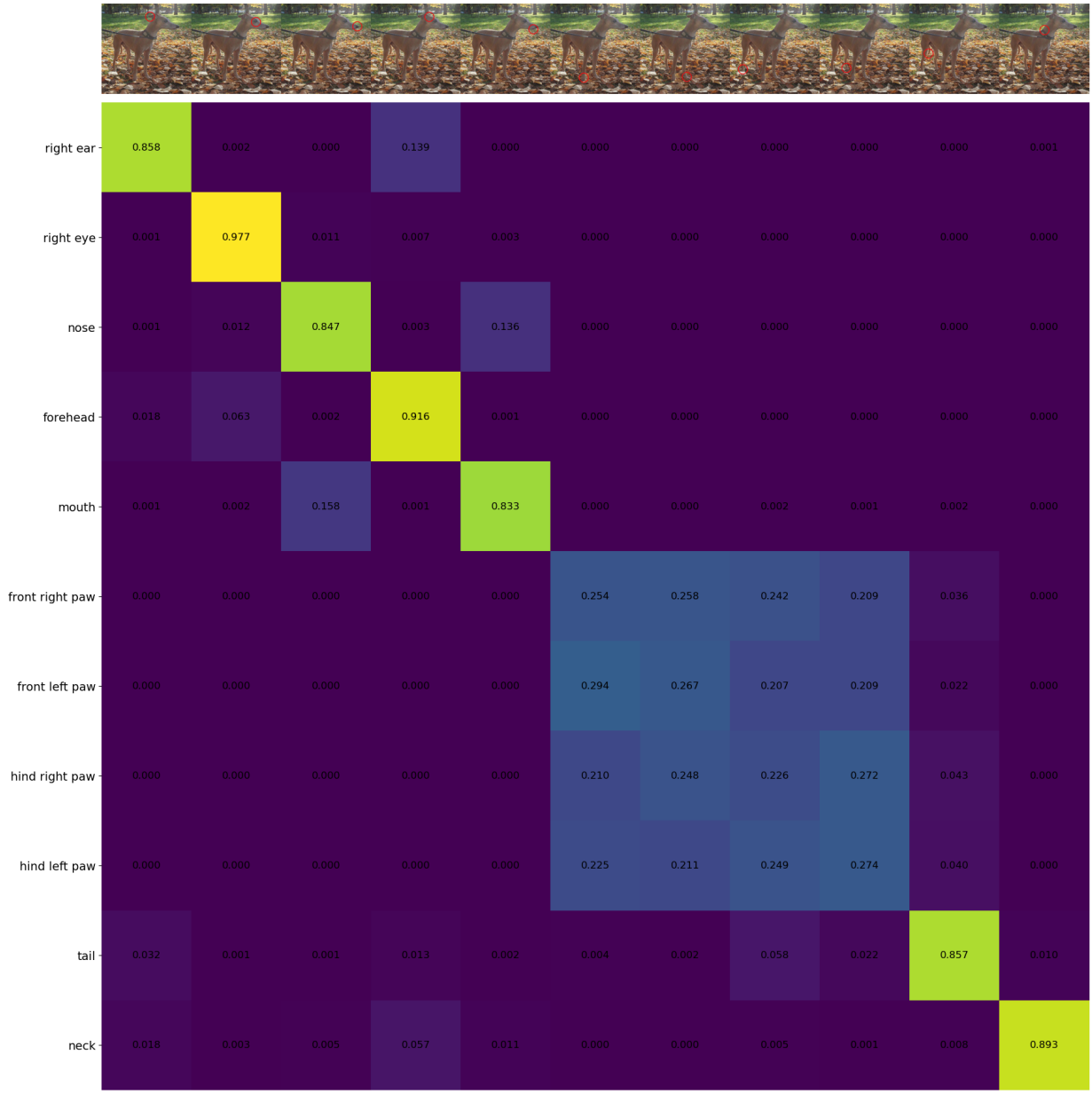

Figure 7: **Naming keypoints.** Normalized cost matrix for an image from SPair71k