Referring Relationships

Ranjay Krishna†, Ines Chami†, Michael Bernstein, Li Fei-Fei
Stanford University
{ranjaykrishna, chami, msb, feifeili}@cs.stanford.edu

1. Supplementary material

In the supplementary material, we include more detailed results of our task for every entity and predicate category, allowing us to diagnose which entities or predicates are difficult to model. We also include the learnt predicate and the inverse predicate shifts for all 70, 4 and 70 predicates we modeled in VRD [3], CLEVR [1] and Visual Genome [2]. Furthermore, we explain our baseline models in more detail here.

Co-occurrence and VRD baseline models

Given that the closest task to referring relationships is referring expression comprehension [4], we draw inspiration from this literature when designing our baselines. A frequent approach used by most models for this task involve semantically mapping language expressions to their corresponding image regions [5, 4, 6]. In other words, they map the image features extracted from a CNN close to the language expression features extracted from a Long Short Term Memory (LSTM). Our baseline models (**co-occurrence** and **VRD**) draws inspiration from this line of work and maps relationships to a semantic feature space and maps them close to the image regions to which they refer to using our attention module.

The difference from the two baseline models is determined by how we embed the relationships to that semantic space. In the case of **co-occurrence**, we are only interested in studying how well we can model relationship without the predicate and rely simply on co-occurrence statistics. So, we first embed the `subject` and the `object`, concatenate their representations and pass them through a dense layer followed by a RELU non-linearity to allow the two embeddings to interact. For the **VRD** baseline, we embed the entire relationship similar to prior work [3] by embeddings all three components of the relationship, concatenating their representation and passing them through a dense and non-linear layer.

Unlike our model, which attends over the subject and object in succession, these models are jointly aware of the entire relationship or at least about the other entity when attending over the image features. Also embedding the predicate and attending over the image with this embedding asks these baselines to model predicates as visual. But predicates such as `above` or `below` are not visually significant and can only be modelled as a relative shift from one entity to another. We show through our experiments that such baselines are not able to perform as well as our model nor are interpretable.

Spatial shift baseline model

Instead of learning the attention shifts for each predicate, we assume (incorrectly) that all predicates are simply spatial shifts and model each predicate as a shift function. We learn the shift statistically from the relative locations of the two entities of the relationship. We visualize these statistically calculated shifts in Figures 3, 5 and 7. We normalize the shifts so visualize the heatmaps. They don't show the actual values of how much each predicate shifts attention but only shows the direction of the shift. We see the as expected `left` push attention to the right, etc. This baseline uses our attention modules to find the subject and object and uses these precalculated shifts to move attention around. We only need to train the attention module, which is equivalent to training our SSAS model with zero iterations. During evaluation, we use these statistical spatial shifts to move attention.

This baseline is useful in two ways. First, it demonstrates that it is important to model predicates as both spatial as well as semantic. Second, it allows us to compare the learnt predicate shifts with these calculated ones to verify that our SSAS models are in fact learning spatial shifts as well.

1.1. Learnt predicate shifts

While `above` and `below` are spatial predicates, others like `hit` or `sleep on` are both spatial as well as semantic. `hit` usually refers to entities around the `subject` and are usually `balls`. Similarly, `sleep on` usually refers to something below the `subject` and typically a `bed` or `couch`. We show the learnt predicate shifts of all the predicates in the three datasets in Figures 2, 4 and 6.

As expected most relationships that are spatial are interpretable. In Figure 2, `above` moves attention below

while its inverse moves it up. hit focuses on the right bottom, emulating the dataset bias of right handed people hitting tennis or baseball. In Figure 6, wearing shifts attention all over the body of the subject focusing mainly on shirts, pants and glasses. By splits the attention both to the left and to the right to find what the subject is next to. Some predicates, like attached to are harder to interpret as they depend on both the semantic as well as spatial shifts. While our model uses the image features to learn these shifts, our current spatial shift visualization does not create an interpretable predicate shift.

1.2. Predicate analysis

One of the benefits of referring relationships is its structured representation of the visual world, allowing us to study which entities and predicates are hard to model. In this section we report the Mean IoU of our model on all the predicate categories for the three datasets in Tables 1 and 3. Note that we don't report the results for CLEVR here since all the 4 spatial predicates are equally represented in the dataset and perform equally across all categories.

Across most predicates we find that the object localization is much harder than the subject's. This occurs because most objects tend to be smaller objects which are better localized by first attending over the subject first. We also see that size is an important factor in detection as predicates like carry and use usually have a larger subject and a smaller object and we find that the IoU for the subject is much higher than that of the object. We also see that when entities are partially occluded, for example <subject - drive - object>, the object IoU is much higher than the occluded subject.

1.3. Object analysis

We run a similar analyze of the performance of our model across all the entity categories and report Mean IoU results in Tables 2 and 4. Note that we don't report the results for CLEVR here since all the entities perform equally across all categories.

We find that the Mean IoU for all entities in Visual Genome are higher than the ones in VRD, implying that more data for each of these categories helps the model learn to attend over the right image regions. In Figure 4, we find that with the predicate shifts, we can detect smaller objects, like face, ear, bowl, eye, a lot better. Some entities like shelves and light don't perform well on the dataset because not all the shelves or light sources are annotated in the dataset, causing the model's correct predictions to be penalized. Surprisingly, the model has a hard time finding bags, perhaps because it learns that bags are often found being worn or carried by people in the training set but the test set contains bags that are on the ground or resting against other entities.

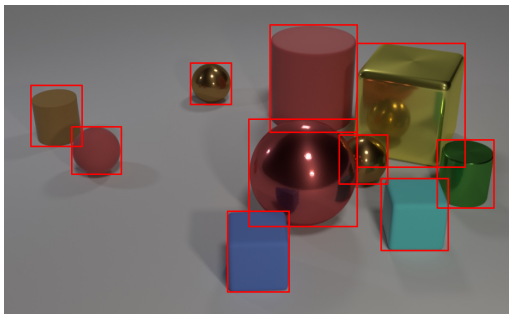

Figure 1: Example bounding box annotations we added to the Clevr dataset.

1.4. CLEVR annotations

The CLEVR dataset is annotated with objects in 3D space [1]. To use the dataset in the same manner as VRD [3] and VisualGenome [2], we converted all the 3D entity locations into 2D bounding boxes, with respect to the viewing perspective of every image. We will release the conversion code as well as the bounding box annotations that we added to CLEVR. Figure 1 showcases an example image annotated with our bounding boxes.

References

[1] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *arXiv preprint arXiv:1612.06890*, 2016. 1, 2

[2] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 1, 2, 5, 6

[3] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pages 852–869. Springer, 2016. 1, 2, 3, 4

[4] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 1

[5] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016. 1

[6] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016. 1

Predicate	S-IoU	O-IoU	Predicate	S-IoU	O-IoU	Predicate	S-IoU	O-IoU
on	0.2904	0.5482	wear	0.4189	0.2830	has	0.4490	0.2339
next to	0.3338	0.3867	outside of	-	0.7778	sit next to	0.3158	0.3152
stand next to	0.4429	0.4436	park next	0.4012	0.5426	sleep on	0.3543	0.5429
above	0.5653	0.4525	behind	0.3055	0.4770	stand behind	0.5748	0.4424
sit behind	0.5854	0.9111	park behind	0.8545	0.5050	in the front of	0.3644	0.4009
under	0.4639	0.5188	stand under	0.2304	0.3622	sit under	0.2716	0.3158
near	0.2964	0.3642	rest on	0.4283	0.4603	walk	0.5814	0.6667
walk past	0.6000	0.8571	in	0.3073	0.4339	below	0.4272	0.5337
beside	0.2939	0.3870	follow	0.4249	0.5367	over	0.5403	0.5055
hold	0.3867	0.1535	by	0.2705	0.4423	beneath	0.4888	0.5282
with	0.3522	0.2823	on the top of	0.2896	0.4416	on the left of	0.2290	0.3272
on the right of	0.2864	0.3338	sit on	0.4281	0.4271	ride	0.4513	0.4936
carry	0.3334	0.1744	look	0.3344	0.2951	stand on	0.3854	0.7179
use	0.4726	0.1160	at	0.2995	0.5185	attach to	0.4193	0.6047
cover	0.3349	0.4364	touch	0.3426	0.4461	watch	0.3022	0.3982
against	0.1364	0.6898	inside	0.1779	0.4751	adjacent to	0.7539	0.6492
across	0.4460	0.5010	contain	0.3174	0.2443	drive	0.1168	0.6528
drive on	0.7723	0.8269	taller than	0.4431	0.4423	eat	0.4726	-
park on	0.4639	0.7347	lying on	0.3457	0.6335	pull	0.4737	0.3362
talk	0.7453	0.1767	lean on	0.5046	0.5127	fly	0.4517	0.2156
face	0.3219	0.5598	play with	0.5735	0.2647			

Table 1: Mean IoU results for referring relationships per predicate in the VRD [3] dataset.

Entity	S-IoU	O-IoU	Entity	S-IoU	O-IoU	Entity	S-IoU	O-IoU
person	0.3909	0.4191	sky	0.7651	0.7602	building	0.3635	0.4707
truck	0.4477	0.5754	bus	0.5864	0.6578	table	0.4693	0.5664
shirt	0.3495	0.3231	chair	0.2103	0.2448	car	0.3293	0.3764
train	0.5213	0.5688	glasses	0.1682	0.2324	tree	0.3106	0.3398
boat	0.2832	0.4775	hat	0.2368	0.2606	trees	0.4637	0.5840
grass	0.5393	0.5474	pants	0.3612	0.3161	road	0.6776	0.6812
motorcycle	0.5031	0.5291	jacket	0.3288	0.3316	monitor	0.3130	0.3404
wheel	0.3348	0.2370	umbrella	0.2670	0.3426	plate	0.2011	0.2899
bike	0.4091	0.3479	clock	0.2273	0.2193	bag	0.0951	0.0915
shoe	-	0.1143	laptop	0.3319	0.3178	desk	0.5790	0.5945
cabinet	0.1700	0.1845	counter	0.3477	0.4249	bench	0.3671	0.4308
shoes	0.2944	0.2879	tower	0.4315	0.5556	bottle	0.1052	0.0809
helmet	0.2834	0.2533	stove	0.2242	0.2941	lamp	0.1467	0.1692
coat	0.2897	0.3203	bed	0.6702	0.6631	dog	0.3619	0.3510
mountain	0.3915	0.4803	horse	0.5253	0.5527	plane	0.3193	0.6164
roof	0.2859	0.2709	skateboard	0.4013	0.3694	traffic light	0.1067	0.0238
bush	0.2328	0.2312	phone	0.0514	0.0671	airplane	0.5333	0.6694
sofa	0.4597	0.5251	cup	0.1423	0.1030	sink	0.2592	0.2119
shelf	0.0583	0.1278	box	0.0442	0.0996	van	0.2144	0.3710
hand	0.1124	0.0413	shorts	0.2423	0.2547	post	0.0941	0.0971
jeans	0.2449	0.3517	cat	0.3629	0.3238	sunglasses	0.3065	0.1535
bowl	0.2226	0.0494	computer	0.2196	0.1676	pillow	0.1321	0.1797
pizza	0.3882	0.3359	basket	0.1330	0.0751	elephant	0.1761	0.4534
kite	0.2463	0.1843	sand	0.9597	0.7765	keyboard	0.2713	0.2421
plant	0.1793	0.1275	can	0.1605	0.2452	vase	0.1575	0.2536
refrigerator	0.1489	0.1949	cart	0.5619	0.5016	skis	0.1761	0.3398
pot	0.1117	0.0450	surfboard	0.2676	0.2227	paper	0.1525	0.0296
mouse	0.1164	0.1029	trash can	0.0324	0.0692	cone	0.1767	0.1813
camera	0.0124	0.1183	ball	0.0595	0.0556	bear	0.3661	0.3441
giraffe	0.5695	0.5949	tie	0.1129	0.1221	luggage	0.4560	0.5042
faucet	0.1704	0.0565	hydrant	0.4108	0.5458	snowboard	0.2798	0.1804
oven	0.4968	0.3169	engine	0.2016	0.1450	watch	-	0.0233
face	0.0873	0.1798	street	0.6986	0.7291	ramp	0.2341	0.4972

Table 2: Mean IoU results for referring relationships per entity category in the VRD [3] dataset.

Predicate	S-IoU	O-IoU	Predicate	S-IoU	O-IoU	Predicate	S-IoU	O-IoU
wearing a	0.5208	0.3946	made of	0.4430	0.3389	on front of	0.2215	0.6592
with a	0.4370	0.1098	WEARING	0.5125	0.3856	above	0.4642	0.4879
carrying	0.4559	0.1555	has an	0.6672	0.0836	covering	0.6003	0.6558
and	0.4192	0.1644	wears	0.5044	0.3542	around	0.4524	0.5527
with	0.4923	0.3324	laying on	0.4557	0.6832	inside	0.2695	0.6084
attached to	0.2627	0.4524	at	0.4473	0.5085	on a	0.3471	0.4978
of a	0.2968	0.5857	hanging on	0.3166	0.4830	near	0.3931	0.4935
OF	0.3320	0.6058	sitting on	0.4301	0.5331	of	0.3215	0.6172
next to	0.3620	0.4949	riding	0.4959	0.4981	under	0.4276	0.5446
over	0.3719	0.5039	behind	0.3798	0.5849	sitting in	0.4025	0.4852
ON	0.3394	0.5508	eating	0.5277	0.4358	to	0.2768	0.5984
in a	0.3580	0.4629	has	0.6183	0.3341	parked on	0.3851	0.5559
covered in	0.5683	0.4607	holding	0.4716	0.3225	for	0.2892	0.4015
playing	0.5863	0.5625	against	0.3765	0.5524	by	0.3368	0.4593
from	0.2940	0.5188	has a	0.5841	0.3016	standing on	0.4715	0.6338
on side of	0.2453	0.5505	in	0.3574	0.5320	wearing	0.4466	0.1613
watching	0.3033	0.4851	walking on	0.4062	0.5990	beside	0.3592	0.5406
below	0.4370	0.5168	IN	0.4005	0.5802	mounted on	0.3054	0.5426
have	0.5750	0.2201	are on	0.3510	0.6001	are in	0.4185	0.6917
in front of	0.3963	0.5210	looking at	0.4503	0.4787	belonging to	0.3250	0.6243
on top of	0.3803	0.5735	holds	0.5194	0.3834	inside of	0.2398	0.3430
along	0.3647	0.5030	hanging from	0.2508	0.2905	standing in	0.4748	0.6173
says	0.1200	-	painted on	0.2632	0.6049	between	0.4090	0.4987

Table 3: Mean IoU results for referring relationships per predicate in Visual Genome [2].

Entity	S-IoU	O-IoU	Entity	S-IoU	O-IoU	Entity	S-IoU	O-IoU
giraffe	0.6361	0.6468	bowl	0.2602	0.3144	food	0.4410	0.4512
face	0.3762	0.4020	people	0.3210	0.3492	shirt	0.3950	0.3774
bench	0.4204	0.5398	light	0.1561	0.1574	head	0.3918	0.4283
zebra	0.6152	0.6127	cow	0.5079	0.5867	sign	0.2390	0.3593
motorcycle	0.5093	0.5648	floor	0.4870	0.5673	hat	0.3891	0.3270
sheep	0.4988	0.4735	truck	0.4420	0.6199	water	0.4134	0.5766
chair	0.2987	0.3421	field	0.6578	0.7331	door	0.2338	0.2906
pizza	0.6415	0.4513	tree	0.3440	0.4077	car	0.3467	0.5168
leg	0.3136	0.3383	bag	0.1794	0.1512	fence	0.4272	0.5216
sidewalk	0.3623	0.4631	girl	0.5544	0.5707	leaves	0.2408	0.2376
jacket	0.4363	0.3913	windows	0.2832	0.2672	road	0.5047	0.5897
glass	0.2339	0.2186	bed	0.4867	0.6171	sand	0.4527	0.6028
trees	0.4799	0.4973	player	0.6028	0.6511	helmet	0.3699	0.3809
man	0.5355	0.5971	grass	0.4306	0.5724	cake	0.4235	0.4622
bear	0.6530	0.6794	hand	0.2257	0.2279	cloud	0.4259	0.3843
street	0.4765	0.5590	ground	0.6269	0.6302	airplane	0.6671	0.7176
mirror	0.2132	0.3290	clock	0.4131	0.4533	plate	0.4529	0.5599
ear	0.3029	0.2670	hair	0.3790	0.4054	window	0.2284	0.2473
boy	0.5793	0.6432	clouds	0.4570	0.4644	handle	0.0671	0.1023
counter	0.3018	0.4660	glasses	0.3164	0.3113	pants	0.4308	0.3939
eye	0.2933	0.2427	pole	0.2374	0.2408	line	0.2265	0.2230
wall	0.3599	0.4230	animal	0.4067	0.5630	shadow	0.3007	0.3013
train	0.6389	0.6494	bike	0.5360	0.5238	boat	0.3467	0.4689
horse	0.5631	0.5964	tail	0.3167	0.3189	nose	0.2959	0.2667
beach	0.6542	0.6755	snow	0.5374	0.5755	elephant	0.6877	0.6409
bottle	0.2039	0.1981	surfboard	0.3388	0.3861	cat	0.6501	0.6796
skateboard	0.4036	0.4373	shorts	0.4454	0.3732	woman	0.5019	0.5392
bird	0.4211	0.5768	sky	0.6741	0.7468	shelf	0.1316	0.1928
tracks	0.3826	0.4737	kite	0.4496	0.3150	umbrella	0.3590	0.4102
guy	0.5813	0.6980	building	0.4169	0.5366	dog	0.5649	0.6532
background	0.5510	0.5531	table	0.3601	0.5719	child	0.4880	0.4252
lady	0.5255	0.6257	plane	0.6689	0.6667	desk	0.3536	0.4990
bus	0.6549	0.7362	wheel	0.2778	0.2744	arm	0.2747	0.2918

Table 4: Mean IoU results for referring relationships per entity category in Visual Genome [2].

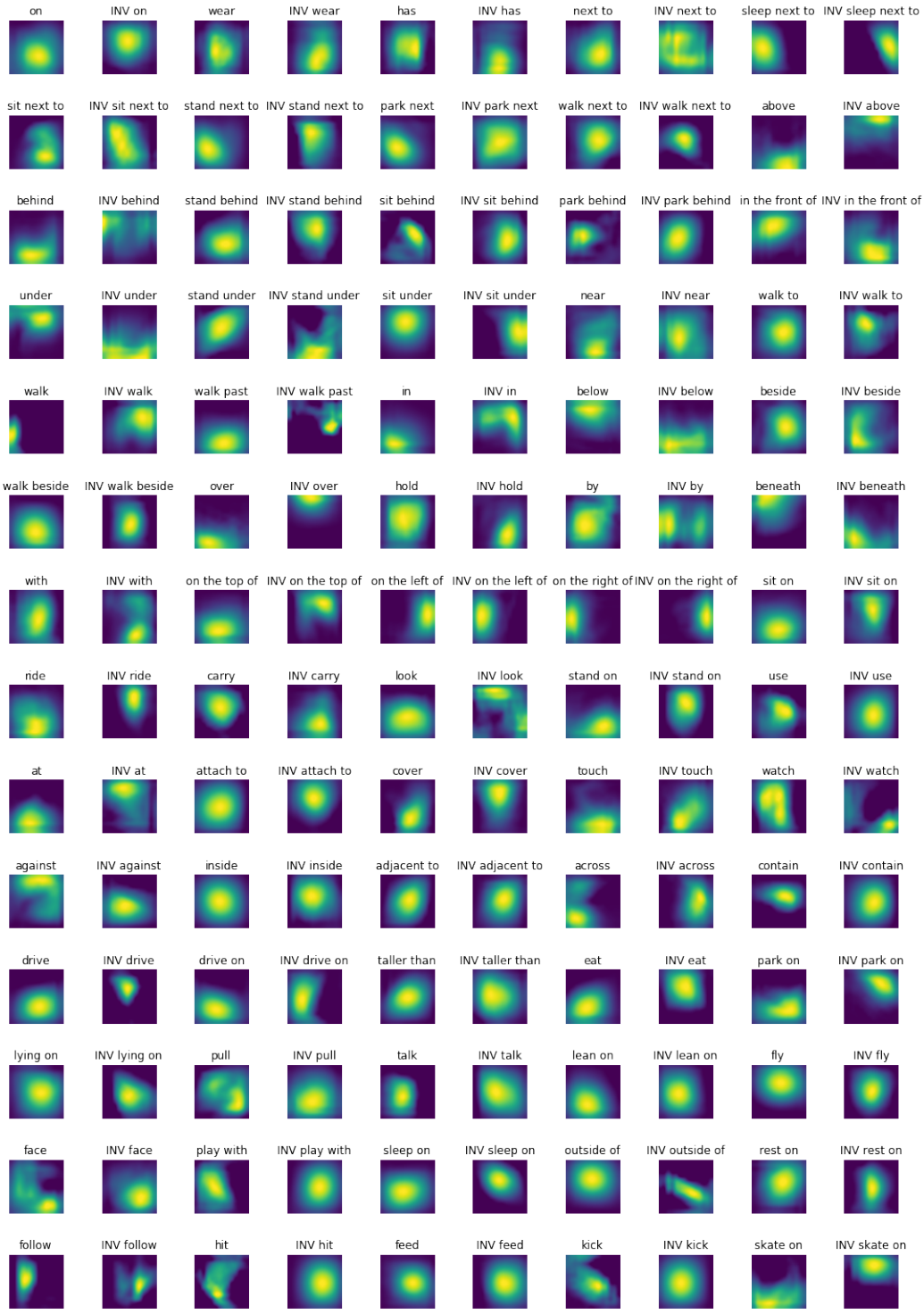

Figure 2: Learnt predicate shifts from the VRD dataset.

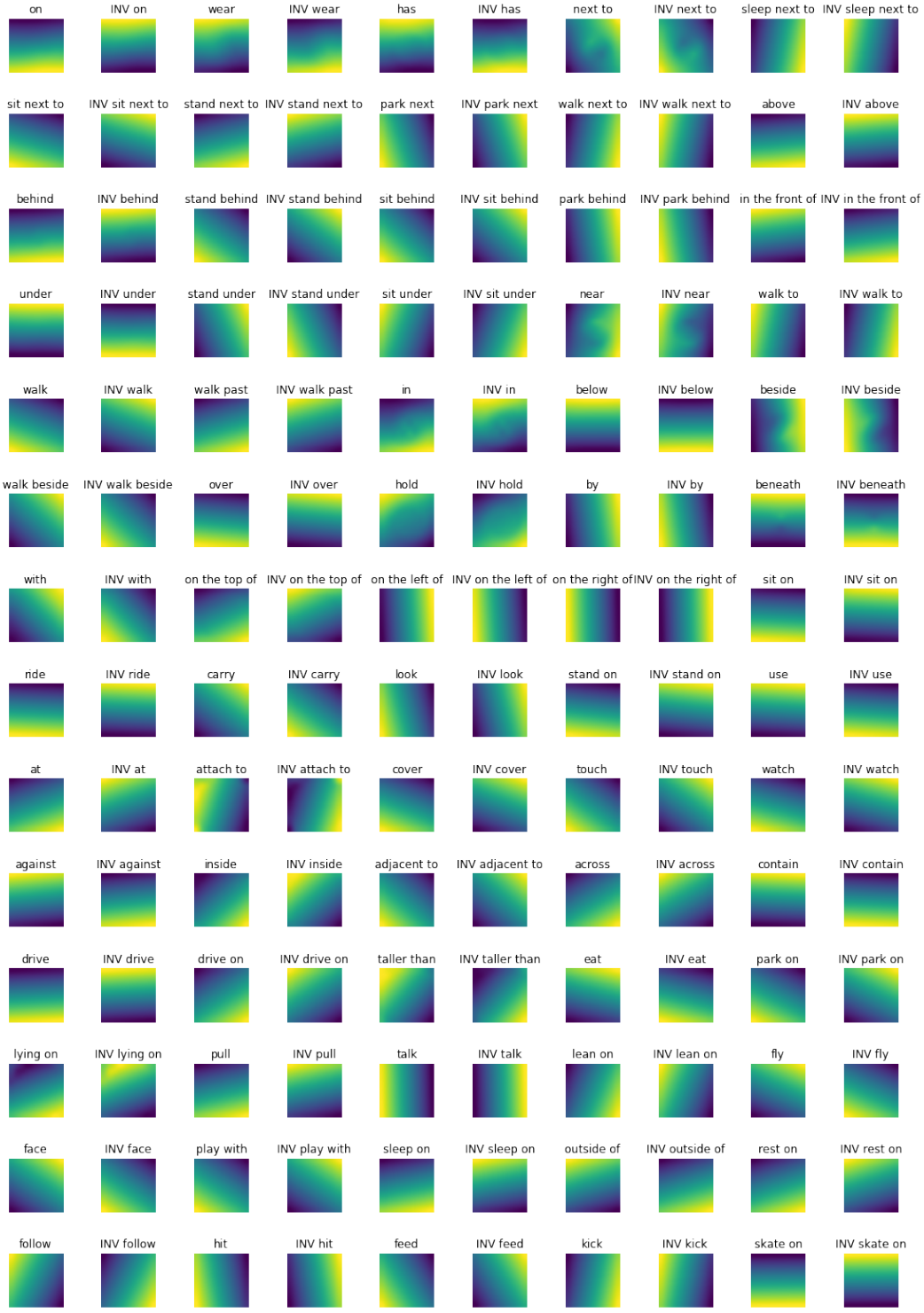

Figure 3: Spatial shifts calculated from the VRD dataset. These shifts were used for the **spatial shift** baseline model.

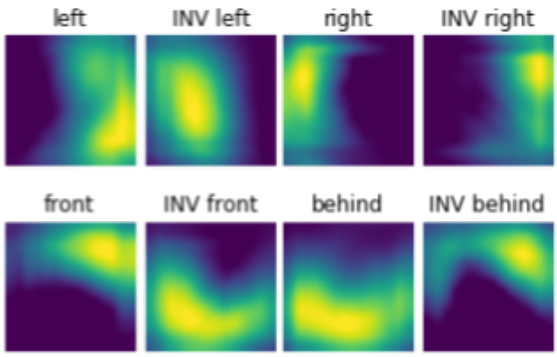

Figure 4: Learnt predicate shifts from the CLEVR dataset.

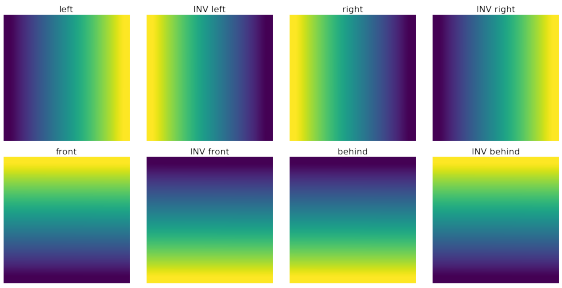

Figure 5: Spatial shifts calculated from the CLEVR dataset. These shifts were used for the **spatial shift** baseline model.

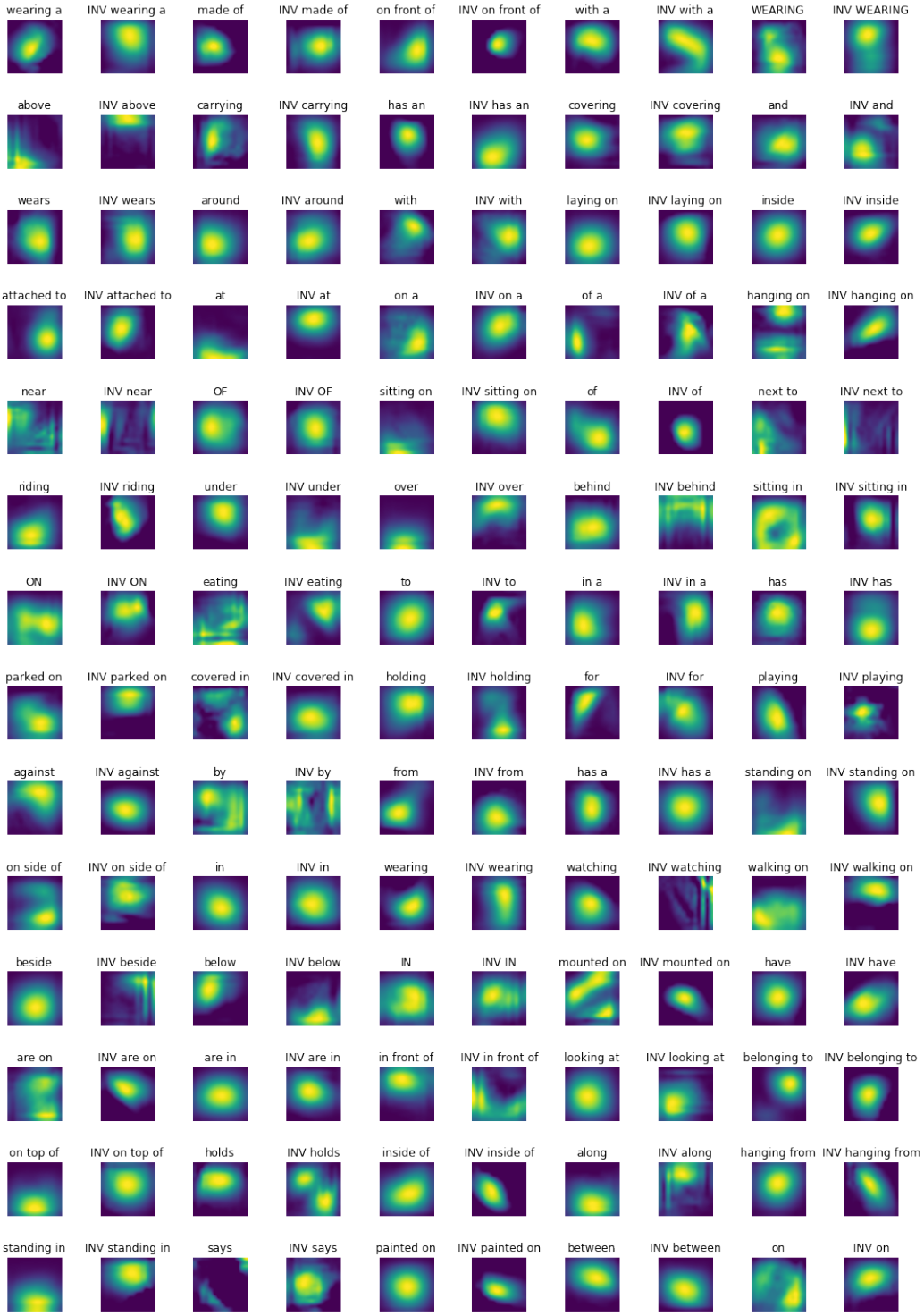

Figure 6: Learnt predicate shifts from the Visual Genome dataset.

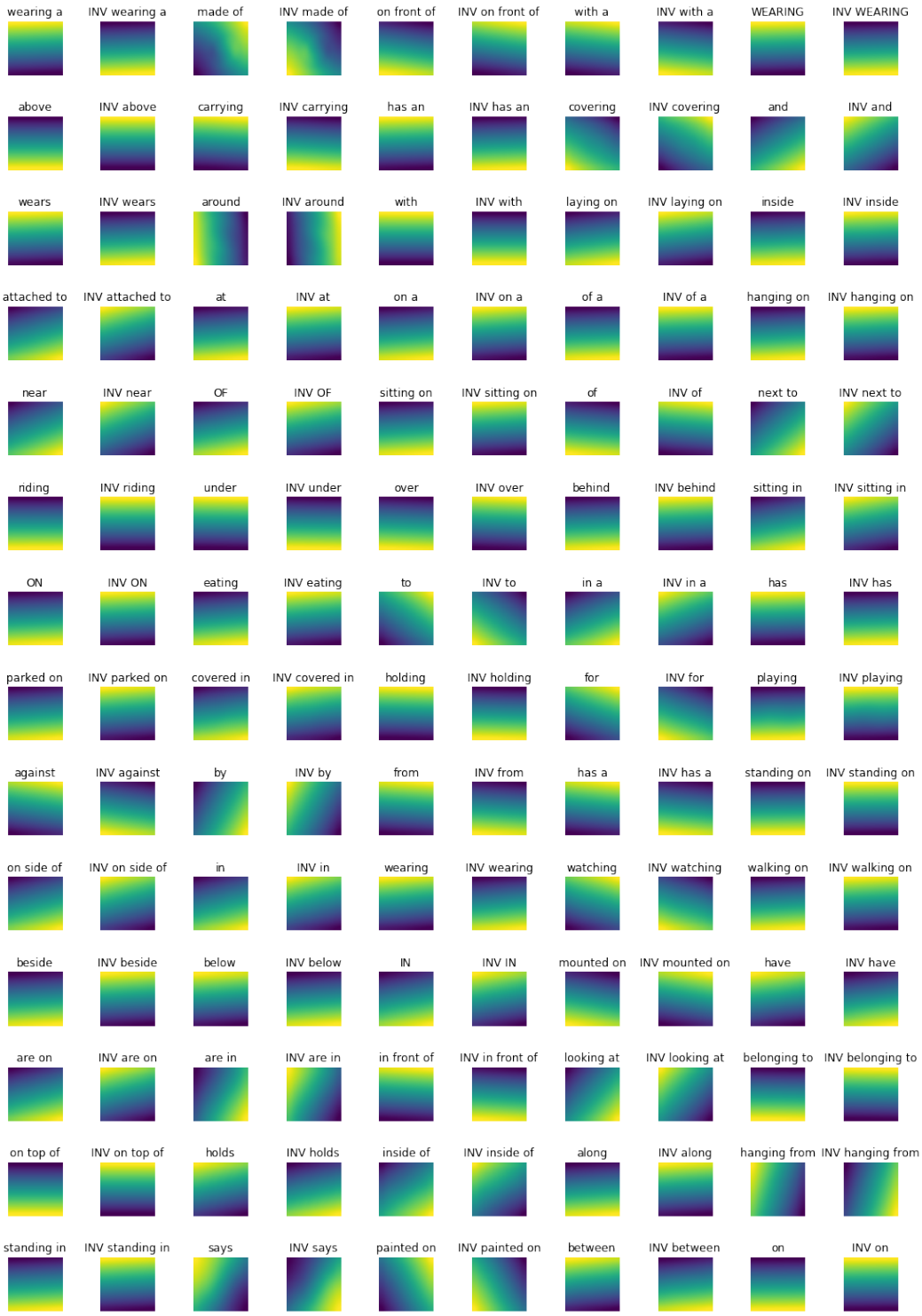

Figure 7: Spatial shifts calculated from the Visual Genome dataset. These shifts were used for the **spatial shift** baseline model.