KG-SP: Knowledge Guided Simple Primitives for Open World Compositional Zero-Shot Learning
– Supplementary Material

Shyamgopal Karthik[1] Massimiliano Mancini[1] Zeynep Akata[1,2]
[1]University of Tübingen [2]Max Planck Institute for Intelligent Systems

1. Further implementation details

We set all hyperparameters of KG-SP following previous works [4, 6] and our MLPs from the implementation of [6], using 3 layers with dimension 768 in the first, 1024 in the second, the number of objects/states in the last, and intermediate Dropout layers with ratio 0.5. For consistency with previous works (*e.g.* [3, 6, 7]), we also test our model without fine-tuning the backbone, denoting it as KG-SP$_{\text{ff}}$ in the tables.

2. Analysis of the ConceptNet Embeddings

In addition to the analysis presented in the main paper, we further analyze the quality of the ConceptNet embeddings both quantitatively and qualitatively.

Comparison with Alternative Word Embeddings. We compare ConceptNet [11] with other popular word embeddings such as Word2Vec [5], Glove [8] and FastText [1], as well as language models [10].

For what concerns the word embeddings, we reject compositions whose cosine similarity between the state and object embeddings is less than 0. For the language models, we use GPT-2 [10] by querying a specific prompt and extracting the likelihood of the next word to be either `Yes` or `No`. We can then set as feasible all compositions whose likelihood of `Yes` is higher than the likelihood of `No`. We tested several prompts (*e.g.* `Can OBJ be STA?`, `You can see a STA OBJ`), but we found the most effective to be: `Question: Is this a STA OBJ? Answer:`, with `STA` and `OBJ` being a state and object respectively.

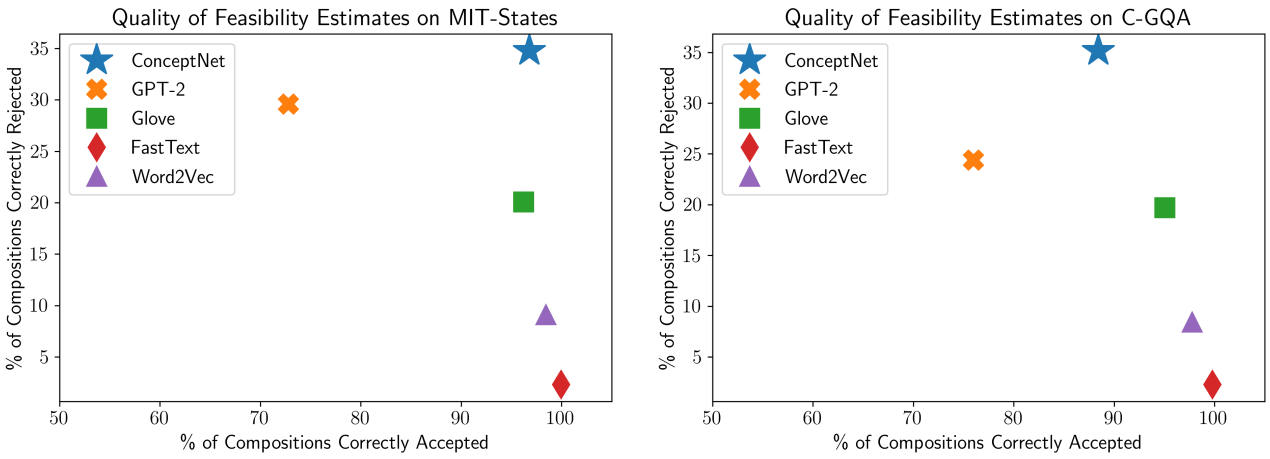
In Fig. 1, we report the results of the feasibility scores computed through the various strategy. Specifically, we report in the x-axis the percentage of compositions correctly considered as feasible (w.r.t. the ones present in the dataset) and on the y-axis the percentage of compositions correctly rejected (*i.e.* compositions not present in the datasets). We report these results for both MIT-states (Fig. 1.a) and C-GQA (Fig. 1.b). In both these datasets, FastText accepts nearly all the compositions therefore, rejecting very few compostions. Similar trends can be seen for Word2Vec and Glove, which reject more unfeasible compositions. While GPT-2 rejects more unfeasible compositions, it also rejects compositions present in the dataset. ConceptNet achieves the best trade-off by accepting most of the compositions present in the dataset, while still rejecting the largest possible number of unfeasible compositions among the competitors.

Feasibility scores	Seen	Unseen	HM	AUC
None	26.3	7.4	7.9	1.3
CompCos (No Tuning)	26.3	7.4	7.9	1.3
CompCos (Tuned)	26.3	7.5	8.0	**1.4**
ConceptNet	**26.5**	**7.7**	**8.2**	**1.4**

Table 1. Comparison of various feasibility scores on the MIT-States for KG-SP$_{\text{ff}}$. We see that despite tuning the threshold for CompCos, ConceptNet is still able to achieve better results.

Quantitative Analysis. In CompCos [4], a method for estimating the feasibility scores of each composition is presented. CompCos computes feasibility scores by using the cosine similarity of learned object/state embeddings, exploiting the available seen compositions during training. Specifically, for a given unseen state-object composition, CompCos computes the cosine similarity of the embeddings of the given object with those of any object containing the same state as seen during training, taking the maximum similarity as feasibility score. The same procedure is applied for the states, and the two scores are averaged.

A natural question is whether our ConceptNet scores can outperform the ones of CompCos. We analyze this in Table 1 for OW-CZSL on the MIT-States validation, using different feasibility scores to perform hard-masking on KG-SP$_{\text{ff}}$, . From the table, we can see that hard masking with the feasibility scores of CompCos does not bring any benefits when used with a natural threshold of zero (*i.e.* half of the range of cosine-similarity scores). When we tune a threshold to



(a) MIT-States (b) C-GQA

Figure 1. Comparison of the ConceptNet embeddings against other word embeddings (Word2Vec, Glove, FastText) and GPT-2. The x-axis denotes the percentage of compositions in the dataset that are correctly accepted. The y-axis denotes the percentage of compositions not present in the dataset that are correctly rejected. We can see that ConceptNet provides an excellent trade-off by rejecting a large number of unfeasible compositions, while still accepting most of the compositions present in the dataset.

filter out less feasible compositions (as in [4]), we see slight improvements (*e.g.* 8.0 vs 7.9 AUC, 1.4 vs 1.3 best HM). However, our ConceptNet-based feasibility scores bring a larger improvement in performance (*e.g.* 7.7 best unseen accuracy, 8.2 best HM), without requiring tuning any threshold or accessing compositional annotations during training. The latter feature makes ConceptNet feasibility scores applicable also in pCZSL, while CompCos-based ones cannot be applied in such setting.

Qualitative Analysis. In this section, we expand the analysis of Section 4.2.1 in the main paper, and we present the top-3 most feasible states and bottom-3 least feasible states for 25 randomly selected objects in Table 2. Similarly, we show the top-3 most feasible objects and the bottom-3 least feasible objects for 25 randomly selected states in Table 3 . As discussed earlier both in Section 4.2.1 in the main paper and in Section 2, ConceptNet provides good estimates of the feasibility of state-object compositions. Overall, the most feasible compositions selected by the embeddings usually turn out to be commonly occurring combinations of the states and objects. For instance, *clear sky*, *burnt flame*, and *engraved jewelry* are all frequently occurring state-object combinations. An interesting aspect of the ConceptNet embeddings is the implicit clustering of the most relevant object/state pairs. For instance, in Table 2 food items (*paste*, *pizza*, *salad*) are linked to cooking-related states states such as *sliced* and *diced*. Similarly, both *chain* and *cord* are linked to the state *frayed*. We can see similar patterns also in Table 3, where *murky* and *muddy* are linked with *mud*, and *unripe* with *fruit*, *pear* and *orange*. However, the estimated feasibility scores are not perfect, and can lead to erroneous outcomes. For instance, this happens for rarely occurring compositions that might be considered unfeasible (*e.g. steaming chains* in Tab. 2, *dirty bracelet* in Tab. 3), and for co-occurring words that might receive high feasibility scores despite not being compatible (*e.g. sunny sea* in Tab. 2, *molten flame* in Tab. 3). This is a limitation of feasibility scores based on single words representations, since these models are biased by the context in which word appear, and their co-occurring frequency. Using a combination of cues from multiple sources may be an effective tool to deal with this issue.

3. Additional Quantitative Experiments

In this section, we report additional experimental results, not included in the main paper due to the lack of space. First, while in the main paper we focused our ablation studies on MIT-States, here we provide additional results also for UT-Zappos and C-GQA. Moreover, we report pCZSL results when applying a bias over the seen compositions, for a direct comparison with OW-CZSL results.

Ablations on other benchmarks. Due to space constraints, in the main paper we performed the ablation studies only on MIT-States, following previous works [3, 4, 6, 9]. However, our results are consistent across settings. As examples, here we also provide ablation studies of the benefit of marginalization on the UT-Zappos dataset in Tab 4. Again, we see that marginalization consistently brings improvements to CGE. We also evaluate the benefit of hard masking on the C-GQA dataset in Tab. 5. We again see hard masking bringing improvements to the HM and AUC on the validation set of C-GQA for both VisProd and KG-SP$_{\text{ff}}$.

Objects	States	
	Most Feasible (Top-3)	Least Feasible (Bottom-3)
balloon	inflated, deflated, filled	grimy, rusty, raw
bear	heavy, large, huge	crinkled, dark, damp
bridge	narrow, curved, bent	barren, cluttered, pressed
building	standing, tall, moldy	ruffled, sharp, pureed
bush	mossy, blunt, wilted	ripped, broken, grimy
cable	coiled, frayed, loose	empty, filled, barren
camera	raw, sharp, lightweight	pierced, crinkled, ruffled
chains	loose, broken, frayed	pureed, unripe, steaming
chair	standing, upright, bent	pierced, thawed, sliced
concrete	unpainted, crushed, molten	ruffled, folded, whipped
cord	coiled, frayed, loose	foggy, dirty, grimy
flower	wilted, ruffled, verdant	grimy, cored, scratched
jewelry	engraved, pierced, worn	steaming, squished, full
mirror	painted, dented, shiny	unripe, pureed, cooked
paste	mashed, pureed, sliced	tall, standing, winding
pizza	sliced, cooked, diced	bright, clear, modern
salad	diced, mashed, sliced	tight, narrow, smooth
sea	sunny, murky, fresh	closed, unpainted, squished
shower	wet, damp, steaming	burnt, cored, thin
snake	coiled, winding, curved	creased, crumpled, pressed
steel	molten, rusty, shiny	runny, unripe, verdant
steps	standing, winding, straight	viscous, tight, dry
stream	muddy, winding, spilled	dented, rusty, caramelized
sugar	caramelized, whipped, melted	scratched, ancient, coiled
tile	painted, unpainted, moldy	whipped, inflated, ripe

Table 2. Examples of top-3 and bottom-3 states associated to 25 randomly selected objects, according to ConceptNet feasibility scores.

States	Objects	
	Most Feasible (Top-3)	Least Feasible (Bottom-3)
ancient	stone, bronze, ceramic	candy, sandwich, cake
bright	lightbulb, sky, orange	sandwich, pizza, beef
browned	sauce, butter, beef	key, vacuum, wall
brushed	coat, wool, dust	cave, lake, building
burnt	flame, fire, smoke	bathroom, shower, camera
clear	sky, concrete, glass	sandwich, pants, pizza
cooked	meat, soup, sauce	key, mirror, moss
crushed	ground, lemon, concrete	lake, pond, clock
curved	blade, steel, knife	oil, eggs, bag
deflated	balloon, bubble, tire	kitchen, wood, coffee
dirty	dirt, mud, dust	balloon, bracelet, cord
dull	blade, knife, sword	fig, mountain, cookie
filled	vacuum, bag, bucket	cable, cliff, sword
heavy	metal, bear, handle	fig, pool, pond
moldy	basement, dust, carpet	tiger, road, highway
molten	metal, flame, copper	book, cat, furniture
murky	cloud, mud, pond	tire, wheel, nut
old	building, roots, tree	sandwich, bubble, chocolate
open	door, window, gate	bus, cliff, granite
peeled	orange, potato, fruit	coffee, tea, tower
thin	paper, blade, paste	garage, car, garden
tiny	penny, toy, town	gear, beef, field
torn	fabric, paper, clothes	pond, bronze, clock
unripe	fruit, pear, orange	mirror, cat, steel
viscous	foam, mud, paste	garden, clock, city

Table 3. Examples of top-3 and bottom-3 objects associated to 25 randomly selected states, according to ConceptNet feasibility scores.

pCZSL: results with bias. In the main paper, we evaluate the models on the pCZSL without applying a bias on the seen compositions. This is because, in the partial setting, we do not have the explicit notion of seen or unseen compositions, thus we cannot follow the same evaluation protocol of standard CZSL and OW-CZSL. Here we show the results on pCZSL when assuming the seen compositions to be known, and applying a bias to them. We underline that this is not fair in the pCZSL setting, but allows us to directly compare results of pCZSL and OW-CZSL. The results are shown in Tab. 6 for all 3 datasets. Consistently with the results without bias, we find that KG-SP always outperforms CompCos and CGE. Specifically, KG-SP achieves significantly better AUC, best harmonic mean, and either comparable or superior best accuracy on seen and unseen compositions. More importantly, the drop from the open-world setting is quite small, *e.g.* 0.78 to 0.61 AUC on C-GQA. This is remarkable, since in pCZSL we have half the labels of OW-CZSL, and indicates the robustness of KG-SP to the available annotations.

	Marginaliz.	Seen	Unseen	HM	AUC
CGE_{ff}		51.3	30.0	25.2	11.5
	✓	**51.4**	**46.6**	**30.0**	**15.4**
CGE		51.1	47.6	**33.3**	17.5
	✓	**53.9**	**48.5**	32.3	**18.1**

Table 4. OW-CZSL results in the validation set of UT-Zappos when using marginalization. CGE_{ff} is the approach of [6] with frozen backbone whereas CGE performs end-to-end training.

	Hard Masking	Seen	Unseen	HM	AUC
VisProd		24.8	14.8	13.2	2.8
	✓	**24.9**	**14.9**	**13.3**	**2.9**
KG-SP		30.5	16.9	15.4	3.9
	✓	**30.6**	**17.0**	**15.5**	**4.0**

Table 5. OW-CZSL results in the validation set of C-GQA when using hard masking.

4. CZSL vs OW-CZSL: an example

In the main text we described the OW-CZSL setting, proposed in [4]. Here, we provide an example to clarify how this setup differs from the more standard CZSL. For more

Method	MIT-States				UT Zappos				C-GQA			
	S	U	HM	AUC	S	U	HM	AUC	S	U	HM	AUC
CompCos [4]	18.2	**6.3**	5.6	0.64	56.0	42.7	34.0	18.4	22.0	1.8	2.9	0.31
CGE [6]	19.2	5.5	5.2	0.61	**60.4**	**43.4**	34.7	19.5	26.7	1.2	2.1	0.25
KG-SP	**20.5**	**6.3**	**5.9**	**0.77**	60.0	43.3	**40.2**	**22.6**	**29.2**	**2.4**	**4.1**	**0.61**

Table 6. **pCZSL results** on MIT-States, UT Zappos and C-GQA. We measure best seen (S) and unseen accuracy (U), best harmonic mean (HM), and area under the curve (AUC) on the compositions.

details on this setting, please refer to [4].

Let us consider a toy benchmark where the training set contains images of the following compositions: *wet cat*, *dry apple*, *dry dog*, *ripe apple*. Similarly, let us assume that the same benchmark contains images of the following unseen compositions in the test set: *wet dog*, *dry cat*. Note that other three compositions, *i.e*. *wet apple*, *ripe dog*, *ripe cat*, are not present in any image of the dataset, either because they are unfeasible (*e.g*. *ripe dog*) or because no image has been collected for them (*i.e*. *wet apple*).

The main difference between CZSL and OW-CZSL is that, in the latter, we have no priors on unseen compositions, and we thus consider the full compositional space at test time. To clarify, in standard CZSL, a model assumes to know which unseen compositions are present in the test set of the dataset and which are not. Thus, a CZSL model would predict 6 compositions: the 4 seen ones during training, and the 2 unseen ones that have at least one image in the test set (*i.e*. *wet dog*, and *dry cat*).

In OW-CZSL, we do not know which compositions are present in the test set. As a consequence, the output space needs to consider all possible unseen compositions. In our toy benchmark, this means predicting the 4 seen compositions and all the 5 compositions for which we did not have training images (*i.e*. *wet dog*, *dry cat*, *wet apple*, *ripe dog*, *ripe cat*). Note that in OW-CZSL a model needs to cope with the presence of "distractors", *i.e*. compositions close to other existing ones but not present in the dataset (*e.g*. *wet apple* vs *ripe apple*) as well as modeling unfeasible compositions (*e.g*. *ripe dog*) to simplify the task. While this is a toy example, the difference between the settings is huge in CZSL benchmarks, where these challenges are more pronounced. As an example, CZSL models on MIT-States consider only 1'662 compositions out of the possible 28'175 compositions considered in the output space of OW-CZSL.

As a final note, despite the difference in the output spaces, models built for standard CZSL may perform well in OW-CZSL since they can still exploit what learned from seen compositions to generalize to unseen ones. This can be seen in Table 1 of the main paper, where CGE [6], designed for standard CZSL, is competitive in most benchmarks, being either third or second best performing model.

Dataset	Licence
UT-Zappos	Custom: Non-commercial Usage
MIT-States	Not Available
C-GQA	Creative Commons Attribution 4.0 License

Table 7. The datasets employed in the paper and their licences.

5. Dataset Licenses

The datasets used in our work are: UT-Zappos [12, 13], MIT-States [2], and C-GQA [6] and their licenses are summarized in Table 7. For MIT-States we did not find an accompanying license, but the dataset is publicly available[1].

References

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 2017. 1

[2] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *CVPR*, 2015. 4

[3] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. In *CVPR*, 2020. 1, 2

[4] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *CVPR*, 2021. 1, 2, 3, 4

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, 2013. 1

[6] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *CVPR*, 2021. 1, 2, 3, 4

[7] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *ECCV*, 2018. 1

[8] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 1

[9] Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In *ICCV*, 2019. 2

[1] http://web.mit.edu/phillipi/Public/states_and_transformations/index.html

[10] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. In *ICML*, 2019. 1

[11] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, 2017. 1

[12] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, 2014. 4

[13] Aron Yu and Kristen Grauman. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *ICCV*, 2017. 4