A. Comparison between GIPCOL and CGE

Although both CGE [18] and *GIPCOL* use GNN to encode compositional concepts, the GNN module functions in a fundamentally different manner in these two models. GNN in *GIPCOL* helps construct the soft prompting for CZSL. However, GNN in CGE plays the text encoder role which projects the concept into the embedding space. *GIPCOLL* freeze CLIP's textual and visual encoders to utilize CLIP's multi-modal aligning ability for CZSL which is more efficient. In contrast, CGE needs to train both the GNN and visual encoder to obtian competitive performance as compared in Fig. 5.

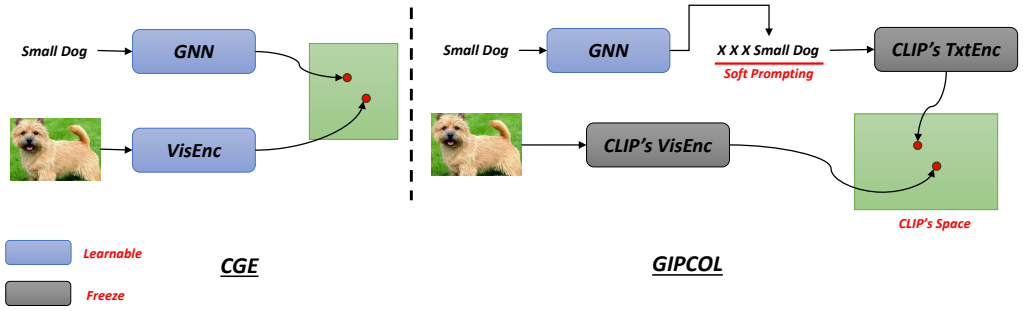

Figure 5. Comparison between CGE and GIPCOL. GIPCOL uses GNN to help prompt construction.

B. GIPCOL Algorithm

Algorithm B.1 *GIPCOL*

1: Initialize GIPCOL using CLIP's pre-trained textual and visual encoders.
2: Update element concept's representation using GNN as Eq. 1 and Eq. 6.
3: Construct textual prompt for compositional labels using the updated element concepts and learnable prefix vectors as Eq. 2.
4: Extract and normalize image/text vectors using CLIP's image/text encoder using Eq. 3 and Eq. 4 seperately.
5: Calculate the class probability as Eq. 5 using the cosine similarity and update GIPCOL's soft-prompting layer Θ and GNN layer Φ using Cross-Entropy loss.

C. CZSL Dataset Statistics

	MIT-States	UT-Zappos	C-GQA
# *Attr.*	115	16	413
# *Obj.*	245	12	674
# *Attr.* $\times$ *Obj.*	28175	192	278362
# Train Pair	1262	83	5592
# Train Img.	30338	22998	26920
# Val. Seen Pair	300	15	1252
# Val. Unseen Pair	300	15	1040
# Val. Img.	10420	3214	7280
# Test Seen Pair	400	18	888
# Test Unseen Pair	400	18	923
# Test Img.	19191	2914	5098

Table 5. Dataset Statistics for MIT-States, UT-Zappos and C-GQA.

D. Feasible Score Threshold in Open-World CZSL

In open-world CZSL (OW-CZSL), we use the validation set to choose a feasible threshold to remove less feasible compositions from the output space and the adopted threshold in *GIPCOL* is shown in Tab. 6.

Dataset	Feasibility Score
MIT-States	0.40691
UT-Zappos	0.51878
C-GQA	0.49941

Table 6. *GIPCOL*'s feasibility threshold score.

E. Qualitative Examples

Figure 6. We show the top-3 predictions of our proposed model for some images. Red colors are ground-truth labels, blue colors are correctly predicted labels and black colors are wrongly predicted labels.

F. Comparison between CLIP's Pre-train Dataset and Target Dataset

We visualize CLIP's pre-training dataset and target domain dataset in Fig. 7. From this figure, we can see that MIT-States have similar visual appearance with CLIP's pre-trained data. However, for UT-Zappos, because of the fashion style change overtime, shoes have significant visual appearance between the pre-training dataset and the target dataset. Results in Tab. 1 and Tab. 2 have shown the domain similarity plays an important role in prompting-based method. Prompting CLIP without any training can achieve better performance on MIT-State then UT-Zappos. *GIPCOL* helps address this challenge partially by prompting design based on the restuls.

Figure 7. Comparison between retrieved images from Laion400M and UT-Zappos/MIT-States.