APPENDIX

Anonymous authors
Paper under double-blind review

A BROADER IMPACT AND LIMITATIONS

Broader Impact This work can be broadly extended to more downstream multi-modality applications, such as general zero-shot learning, text-image retrieval, text-to-image generation, etc., when the class composition is not especially taken into consideration. Besides, the central idea of LLM-grounded modality alignment is not limited to text and image, but any modality that could reveal the semantic categories in practice is promising to explore in the future. The potential negative societal impact is that, the developers should be cautious by carefully examining the societal biases indicated by the generated textual class descriptions, even though the large language models we used are publicly accessible.

Limitations One limitation is that the primitive decomposition could be difficult to learn when the states are non-visual concepts like `smelly`, `hot`, etc., even by the pre-trained CLIP model. Another limitation is that the generated descriptions by LLMs are not grounded to the image such that some distraction from generated descriptions could be introduced.

B GENERATING COMPOSITIONAL CLASS DESCRIPTIONS

In this work, we choose T5 and OPT models as the LLMs for compositional class description generation. For the T5 model, we follow the same setting as (He et al., 2023) that uses the T5-base model for word-to-sentence generation. The T5-base model was pre-trained on the Colossal Clean Crawled Corpus dataset (Raffel et al., 2020) and finetuned on the CommonGen dataset (Lin et al., 2020). Take the `painted ceiling` as an example, the results from T5-base model are:

```
- A very old but beautifully decorated ceiling.
- A remodeled interior with a painted ceiling.
- A painted ceiling at a restaurant.
- Stained glass windows and a carved pattern on the ceiling.
- Painted ceilings and a fireplace.
- This apartment has a painted ceiling.
- A painted ceiling was a huge hit.
- A chandelier is painted in the middle of the ceiling.
- A stained glass window in a bathroom with a white painted ceiling.
- The ceiling of a hotel is painted.
- ...
```

For the OPT model, we adopt the pre-trained OPT-1.3B model which is small but still could generate sentences of sufficiently good quality. Considering that it is a decoder-only LLM that relies on a few examples as the prompt for word-to-sentence generation, we design the prompt template with 4 examples as shown below.

```
Keywords: sliced, potato, picture
Output: The picture features a beautifully arranged plate of thinly
    sliced potatoes.
###
Keywords: red, apple, photo
Output: In the photo, a bright red apple is the central focus,
    captured in stunning detail.
###
Keywords: leather, shoes, image
Output: The image showcases a sleek pair of leather shoes,
    meticulously designed to impress.
```

```
###
Keywords: broken, car, photo
Output: The photo captures the aftermath of a car accident with a
    broken vehicle lying on the road.
###
```

where the `Keywords` is followed by the words of the state, object, and the word randomly picked from the set {`photo`, `image`, `picture`}. Using the same example *painted ceiling* as T5-base model, the generated sentences are:

```
- The painting of the ceiling features an intricate pattern of
    intricate gold-colored paint and is framed by a white
    background.
- The ceiling has been painted with the pictures of these three
    characters, all arranged together.
- In the picture, the ceiling is covered in bright, colorful
    paintwork that has been painted on by hands painted white. The
    colors have been selected carefully.
- In the picture, the ceiling features painted decoration. The
    decoration resembles the surface of the sea, and has been
    painted in shades of blue.
- The photograph captures both the bright colors of the painting
    atop the ceiling and the subtle shades of light reflecting off
    of it.
- The large picture shows a large pattern painted onto the ceiling.
    The blue line shows paint dripping down.
- The wall behind the picture shows three different painted
    ceilings, in bright contrasting colors. A vibrant sky and blue
    skies are depicted against the dark brick wall.
- The ceiling of the room depicted in the painting could very well
    be painted in a few hours. The details of each object are
    clearly defined in its placement and position.
- Another photo of the same scene, this time featuring a ceiling
    painted in a stunning, white color.
- A painted ceiling is shown, painted according to a specific
    design. this is a typical design that can also include
    decorative or functional elements.
- ...
```

It is clear that the generated class descriptions are much more diverse and informative than those of the OPT model.

C COVARIANCE SHARING

For the CZSL task, the spatial complexity of computing the covariance matrix $\mathbf{\Sigma}_{1:C}$ is $O(|C^{(s)}|^2 d)$ which could be too heavy to compute if the number of the compositions is too large. For example, the C-GQA dataset contains 278K seen compositions which result in around 6×10^{13} floating elements of $\mathbf{\Sigma}_{1:C}$ for 768-dim text features. To handle this issue, we instead implement the $\mathbf{\Sigma}_{1:C}$ by sharing the covariance across attributes given the same object. This implies that the model is encouraged to learn the object-level distributions.

Specifically, similar to the VLPD module of the main paper, we compute the mean $\boldsymbol{\mu}_{1:|\mathcal{O}|}$ and covariance $\mathbf{\Sigma}_{1:|\mathcal{O}|}$ over the objects by grouping $\mathbf{t}_y$ and $\mathbf{D}^{(y)}$ with object labels:

$$\mathbf{t}_o = \frac{1}{|\mathcal{Y}_o|} \sum_{y \in \mathcal{Y}_o} \mathbf{t}_y, \quad \mathbf{D}^{(o)} = \frac{1}{|\mathcal{Y}_o|} \sum_{y \in \mathcal{Y}_o} \mathbf{D}^{(y)}, \tag{1}$$

where $\mathcal{Y}_o$ is the subset of compositions in $\mathcal{Y}$ that contains the same object as y. Then, all the pairwise margins $\mathbf{H}_o^{(m)} \in \mathbb{R}^{|\mathcal{O}| \times |\mathcal{O}|}$ in object space can be mapped back to $\mathbf{H}^{(m)} \in \mathbb{R}^{C \times C}$ in a compositional space by sharing it with all compositions in $\mathcal{Y}_o$. This could significantly reduce the computation load of the covariance while compromising the accuracy of distribution modeling.

Variants	Mem.(GB)	H_{cw}	AUC_{cw}	H_{ow}	AUC_{ow}
ProDA (Lu et al., 2022)	32.5	32.71	16.11	17.30	5.11
PLID (w. ShareCov)	**17.6**	38.50 (-0.47%)	21.69 (-0.43%)	19.81 (-0.60%)	7.04 (-0.30%)
$\mathbb{PLID}$ (full)	22.2	**38.97**	**22.12**	**20.41**	**7.34**

Table 1: Effect of covariance sharing on MIT-States dataset. All methods use the same batch size of 64 for a fair comparison of GPU memory.

Since the distribution modeling for both our $\mathbb{PLID}$ and ProDA is not applicable to the C-GQA dataset, we use the MIT States dataset to show the negative impact of sharing the covariance (see Table 1). It shows that the covariance sharing can significantly save the GPU memory (17.6 vs 32.5 GB), while still performing much better than ProDA.

D Primitive-level Gaussian Modeling

To formulate the Gaussian distributions over the state classes and the object classes, we group the text embeddings of composition descriptions $\mathbf{D}$ by Eq. (1), resulting in the distribution support points (DSP) $\mathbf{t}_o + \mathbf{D}^{(o)}$ and $\mathbf{t}_s + \mathbf{D}^{(s)}$ for a given object class o and state class s, respectively. The DSPs are assumed to follow the state distribution $\mathcal{N}(\mathbf{t}_s, \mathbf{\Sigma}_s)$ or the object distribution $\mathcal{N}(\mathbf{t}_o, \mathbf{\Sigma}_o)$, where the covariances $\mathbf{\Sigma}_s$ and $\mathbf{\Sigma}_o$ are determined by $\mathbf{D}^{(s)}$ and $\mathbf{D}^{(o)}$, respectively.

Eventually, given the decomposed state visual features $f_s(\mathbf{v})$ and object visual features $f_o(\mathbf{v})$, the logit margin terms are defined as

$$h_{k,s}^{(m)} = f_s(\mathbf{v})^\top \mathbf{A}_{k,s} f_s(\mathbf{v}), \quad \text{and} \quad h_{k,o}^{(m)} = f_o(\mathbf{v})^\top \mathbf{A}_{k,o} f_o(\mathbf{v}), \tag{2}$$

where the index k ranges within $[1, |\mathcal{S}|]$ for computing the state classification loss $\mathcal{L}_s$, and ranges within $[1, |\mathcal{O}|]$ for computing the object classification loss $\mathcal{L}_o$, respectively.

E More Implementation Details

Datasets We perform experiments on three CZSL datasets, *i.e.*, MIT-States (Isola et al., 2015), UT-Zappos (Yu & Grauman, 2014), and C-GQA (Naeem et al., 2021). MIT-States consists of 115 states and 245 objects, with 53,753 images in total. Following (Purushwalkam et al., 2019; Nayak et al., 2023; Lu et al., 2023), it is split into 1,262 seen and 300/400 unseen compositions for training and validation/testing, respectively. UT-Zappos contains 16 states and 12 objects for 50,025 images in total, and it is split into 83 seen and 15/18 unseen compositions for training and validation/testing. C-GQA contains 453 states and 870 objects for 39,298 images, and it is split into 5,592 seen and 1,040/923 unseen compositions for training and validation/testing, respectively, resulting in 7,555 and 278,362 target compositions in closed- and open-world settings.

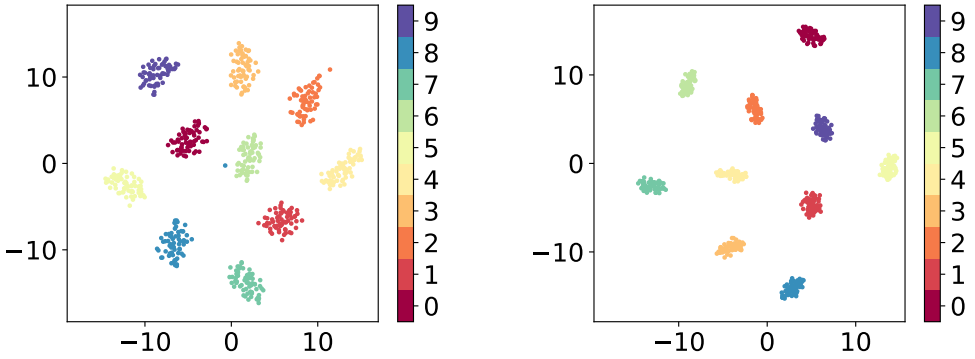
Implementation Our model is implemented on top of the CSP (Nayak et al., 2023) codebase, which extends the CLIP model for compositional zero-shot learning. To tokenize the generated long sentences of each compositional class, we set the context length to the default value of 77 in the original CLIP model. For the soft prompt embeddings, we set the context length of text encoder to 8 for all datasets. We use the dropout rate of 0.3 for the learnable state and object embeddings. In training, we follow the DFSP (Lu et al., 2023) that uses the performance of the validation set for model selection. The rest hyperparameters of our final model on each dataset are listed in Table 2.

F More Results

Primitive-level Visualization In addition to the tSNE visualization of Gaussian distributions over the composition-level classes, we provide the visualizations of the primitive-level classes in Fig. 1. These figures show that our model could learn better text distributions over state classes and object classes than those of the pre-trained LLMs.

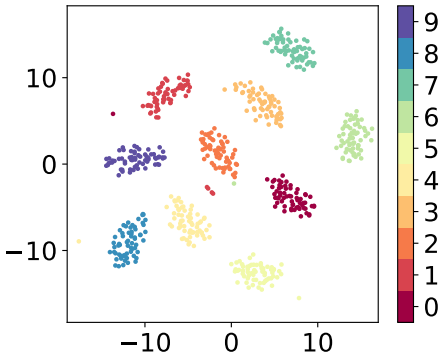
Hyperparameters	MiT-States	UT-Zappos	C-GQA
max epochs	50	25	20
base learning rate	0.00005	0.0001	0.00001
weight decay	0.00002	0.00001	0.00001
number of text descriptions	64	32	64
number of image views	8	8	8
attention dropout	0.5	0.1	0.1
weights of primitive loss	0.1	0.01	0.01

Table 2: Hyperparameters of model implementation.

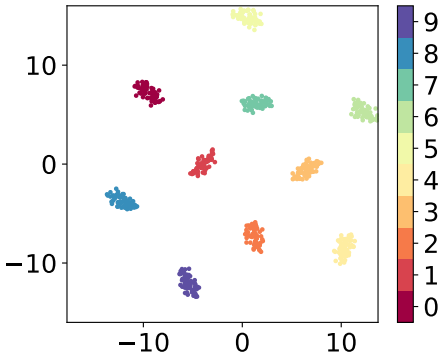


(a) LLM Embeddings of States

(b) Learned DSP of States



(c) LLM Embeddings of Objects



(d) Learned DSP of Objects

Figure 1: tSNE visualization of the primitive-level text embeddings (*states*: Fig. 1a and 1b, *objects*: Fig. 1c and 1d). This figure clearly shows that, compared to the raw embeddings by pre-trained LLMs, our method achieves better distributions over both the state and object classes.

REFERENCES

Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? In *ICLR*, 2023.

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

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *EMNLP*, 2020.

Xiaocheng Lu, Ziming Liu, Song Guo, and Jingcai Guo. Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. In *CVPR*, 2023.

Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *CVPR*, 2022.

Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *CVPR*, 2021.

Nihal V Nayak, Peilin Yu, and Stephen H Bach. Learning to compose soft prompts for compositional zero-shot learning. In *ICLR*, 2023.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(1):5485–5551, 2020.

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