A. Dataset statistics

Fig. 4 shows the distribution of question types in the Super-CLEVR dataset. The question type is determined by the type of the last operation in the question program.

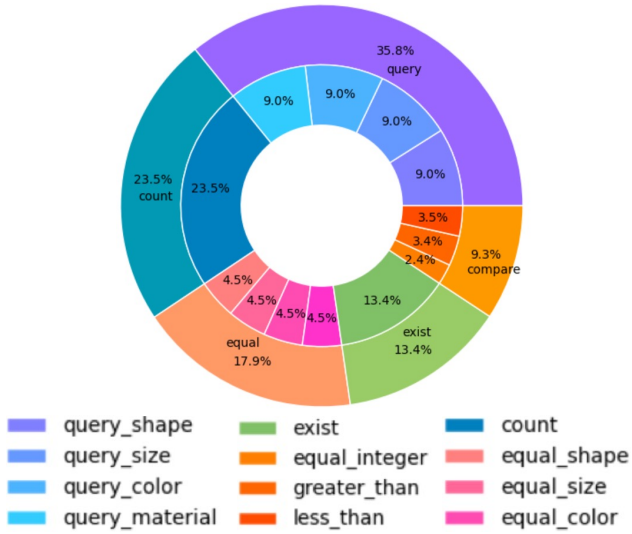


Figure 4. Distribution of question types in Super-CLEVR.

B. List of objects

Super-CLEVR contains 21 objects from 5 categories: airplane, bicycle, bus, car and motorcycle. They are shown in Fig. 5 and Tab. 4.

catetory	objects
airplane	airliner, biplane, jet, fighter
bicycle	utility bike, tandem bike, road bike, mountain bike
bus	articulated bus, double bus, regualr bus, school bus
car	truck, suv, minivan, sedan,wagon
motorcycle	chopper, scooter, cruiser, dirtbike

Table 4. Super-CLEVR dataset contains 21 objects from 5 categories.

C. Dataset controlling

Fig. 6 shows the concept distribution for dataset variants *bal*, *slt* and *long*; and variants *head*, *tail* and *oppo* for testing purpose. Fig. 7 shows the concept co-distribution Matrix M for controlling the concept compositionality (for variants *co-0*, *co-1* and *co-2*). The descriptions for the variants are in Sec. 3.3.

D. Definition of Relative Degrade

Here we describe *Relative Degrade* for each domain shift factors. We use A^i_j to denote the accuracy of model trained using data variant i and tested with data variant j.

Visual complexity:

$$RD = Avg(\frac{A^{easy}_{easy} - A^{easy}_{mid}}{A^{easy}_{easy}}, \frac{A^{easy}_{easy} - A^{easy}_{hard}}{A^{easy}_{easy}}, \frac{A^{mid}_{mid} - A^{mid}_{easy}}{A^{mid}_{mid}}, \frac{A^{mid}_{mid} - A^{mid}_{hard}}{A^{mid}_{mid}}, \frac{A^{hard}_{hard} - A^{hard}_{easy}}{A^{hard}_{hard}}, \frac{A^{hard}_{hard} - A^{hard}_{mid}}{A^{hard}_{hard}})$$

Question redundancy,

$$RD = Avg(\frac{A^{rd-}_{rd-} - A^{rd-}_{rd}}{A^{rd-}_{rd-}}, \frac{A^{rd-}_{rd-} - A^{rd-}_{rd+}}{A^{rd-}_{rd-}}, \frac{A^{rd}_{rd} - A^{rd}_{rd-}}{A^{rd}_{rd}}, \frac{A^{rd}_{rd} - A^{rd}_{rd+}}{A^{rd}_{rd}}, \frac{A^{rd+}_{rd+} - A^{rd+}_{rd-}}{A^{rd+}_{rd+}}, \frac{A^{rd+}_{rd+} - A^{rd+}_{rd}}{A^{rd+}_{rd+}})$$

Concept distribution,

$$RD = \frac{1}{3}\sum_{k\in S}\frac{(A^k_{head} - A^k_{tail}) + (A^k_{long} - A^k_{oppo})}{2 \cdot A^k_k}, S = \{bal, slt, long\}$$

Concept compositionality,

$$RD = Avg(\frac{A^{co\text{-}0}_{co\text{-}0} - A^{co\text{-}0}_{co\text{-}1}}{A^{co\text{-}0}_{co\text{-}0}}, \frac{A^{co\text{-}0}_{co\text{-}0} - A^{co\text{-}0}_{co\text{-}2}}{A^{co\text{-}0}_{co\text{-}0}}, \frac{A^{co\text{-}1}_{co\text{-}1} - A^{co\text{-}1}_{co\text{-}0}}{A^{co\text{-}1}_{co\text{-}1}}, \frac{A^{co\text{-}1}_{co\text{-}1} - A^{co\text{-}1}_{co\text{-}2}}{A^{co\text{-}1}_{co\text{-}1}}, \frac{A^{co\text{-}2}_{co\text{-}2} - A^{co\text{-}2}_{co\text{-}0}}{A^{co\text{-}2}_{co\text{-}2}}, \frac{A^{co\text{-}2}_{co\text{-}2} - A^{co\text{-}2}_{co\text{-}1}}{A^{co\text{-}2}_{co\text{-}2}})$$

E. More details about P-NSVQA

Given a image containing n objects, we maintain a vector of probability $\mathbf{p} = [p^1, p^2, \ldots, p^n]$, where p^k means the probability that object k is selected. We update $\mathbf{p}$ when executing the reasoning operations step by step. In the following, we describe how to compute $\mathbf{p}$ for each kind of operations.

- *scene*
  Initialize all the values in $\mathbf{p}$ to 1.

- $filter_{identifier}[attribute]$ (*e.g.* $filter_{color}[red]$)
  For object k,
  $$p^k = p^k * P^k_{attribute}$$
  Here $P^k_{attribute}$ is the probability of object k having the $attribute$, which is predicted by the visual scene parsing model.

- $relate\_spacial$ (including $relate\_behind$, $relate\_front$, $relate\_right$, $relate\_left$)
  The output of the *relate* operation is the probabilities of each object being on the $spacial$ side of the given object. For example, $relate\_left(i)$ computes the probabilities of objects to be on the left side of the given object i.
  $$p^k_{front} = \frac{1}{1 + e^{-b[(y_k - y_i) + a]}}$$
  $$p^k_{behind} = \frac{1}{1 + e^{-b[(y_i - y_k) + a]}}$$
  $$p^k_{right} = \frac{1}{1 + e^{-b[(x_k - x_i) + a]}}$$
  $$p^k_{left} = \frac{1}{1 + e^{-b[(x_i - x_k) + a]}}$$
  Here i is the input object and (x_i, y_i) is the center of it. a and b are hyperparameters. In our experiments, we set $a = 20$ and $b = 0.02$.

- $same\_color$, $same\_shape$, $same\_size$, $same\_material$
  The *same* operation returns the probabilities of each object having the same attribute as the given object i. For example, for object k and attribute *color*,
  $$p^k = cosine\_similarity(P^k_{color}, P^i_{color})$$

- $intersect$, $union$
  Given two probability vectors $\mathbf{p_1}, \mathbf{p_2}$, we calculate their intersection or union:

$$Intersection : \mathbf{p} = \mathbf{p_1} \odot \mathbf{p_2}$$

$$Union : \mathbf{p} = 1 - (1 - \mathbf{p_1}) \odot (1 - \mathbf{p_2})$$

  Here $\odot$ is the pointwise product.

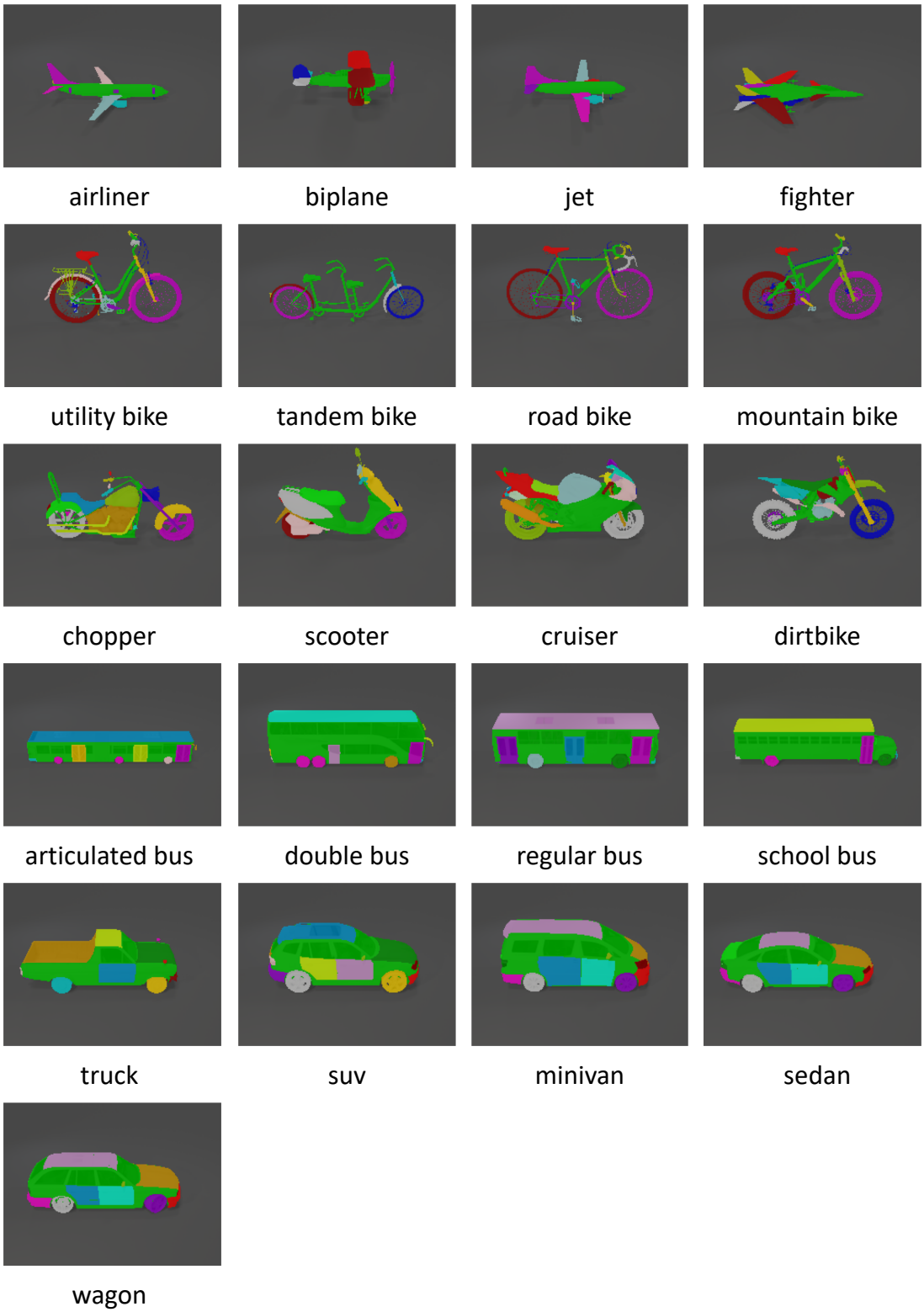


Figure 5. There are 21 objects belonging to 5 categories in the Super-CLEVR dataset.

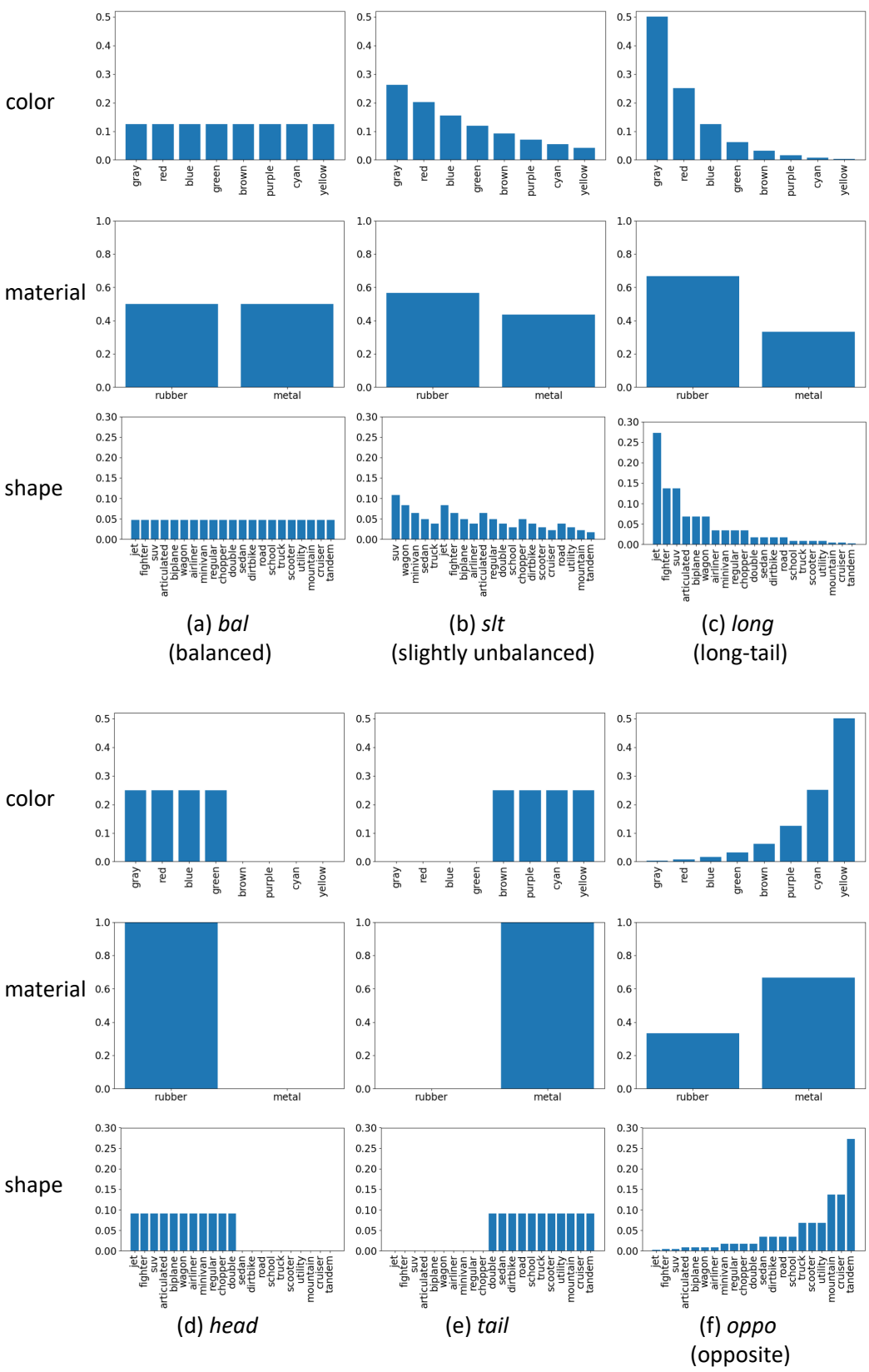
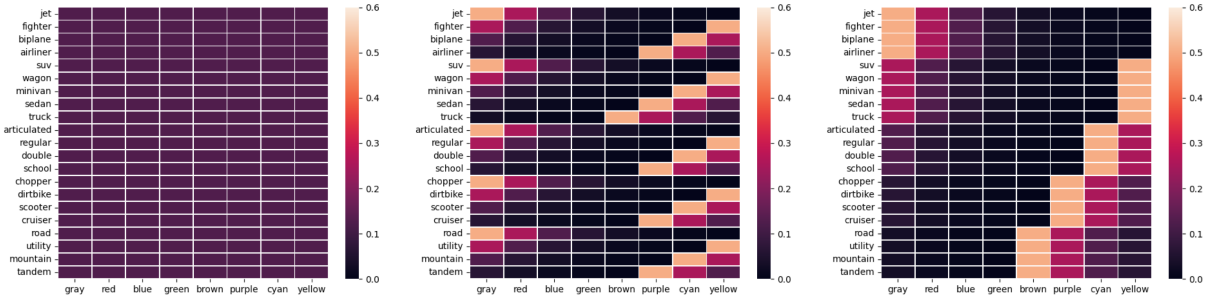


Figure 6. Concept distribution for dataset variants *bal*, *slt*, *long*, *head*, *tail* and *oppo*.



(a) *co-0* (b) *co-1* (c) *co-2*

Figure 7. Concept co-distribution matrix M for dataset variants *co-0*, *co-1* and *co-2*.