Supplementary Materials for: Zero-Shot 3D Shape Correspondence

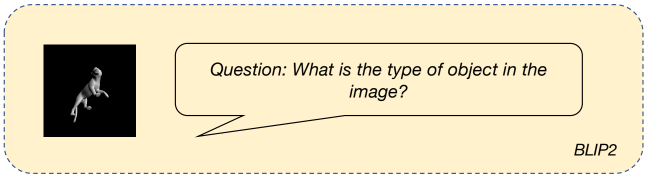

Question: What is the type of object in the image?

Figure 1: The textual prompt for proposing a class label given a rendered image using BLIP2 model.

A IMPLEMENTATION DETAILS

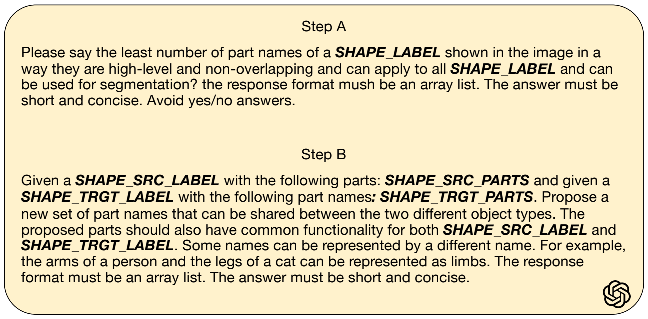
We run all our experiments on a single Nvidia RTX 3090 (24 GB RAM). We use the ChatGPT-3.5 turbo model via OpenAI Python API. We use the Nvidia Kaolin library [Fuji Tsang et al. 2022] written in PyTorch for rendering shapes. We render the mesh on a black background with 512 × 512 resolution. We use a bounding box prediction threshold of 3.7 for the DINO [Caron et al. 2021] model. To ensure fairness, we use the same number of views when comparing SAM-3D and SATR.

B SEMANTIC REGION GENERATION AND MATCHING PROMPTS

In Figure 4, we show the textual prompt we use for proposing sets of semantic regions R^1, R^2 for the input shapes S^1 and S^2 as discussed in Section 3.2. We replace the "SHAPE_SRC_LABEL" and "SHAPE_TRGT_LABEL" strings with the predicted class label for S^1 and S^2, respectively.

C PROMPT CONSTRUCTION TRIALS

We investigated different prompts for obtaining the coarse shape correspondences. First, we try a two-step approach. For each shape separately, we ask Visual-ChatGPT [Wu et al. 2023b] to propose a list of semantic regions given at one time in a rendered image. The answers are then unified using ChatGPT in a similar approach as in Figure 3. Then, we ask ChatGPT to provide a set of semantic regions that can be shared/used for both shapes. We used the prompts, which are shown in Figure 2:

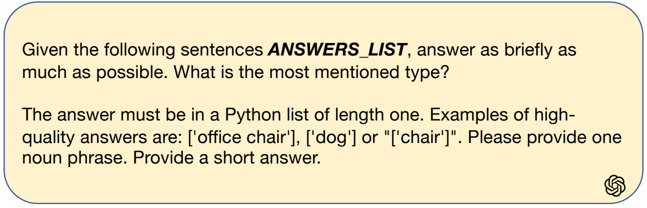

Step A

Please say the least number of part names of a ***SHAPE_LABEL*** shown in the image in a way they are high-level and non-overlapping and can apply to all ***SHAPE_LABEL*** and can be used for segmentation? the response format mush be an array list. The answer must be short and concise. Avoid yes/no answers.

Step B

Given a ***SHAPE_SRC_LABEL*** with the following parts: ***SHAPE_SRC_PARTS*** and given a ***SHAPE_TRGT_LABEL*** with the following part names***:*** ***SHAPE_TRGT_PARTS***. Propose a new set of part names that can be shared between the two different object types. The proposed parts should also have common functionality for both ***SHAPE_SRC_LABEL*** and ***SHAPE_TRGT_LABEL***. Some names can be represented by a different name. For example, the arms of a person and the legs of a cat can be represented as limbs. The response format must be an array list. The answer must be short and concise.

Figure 2: The two-step textual prompt we used for proposing the semantic region per shape and the semantic region mapping.


Given the following sentences ***ANSWERS_LIST***, answer as briefly as much as possible. What is the most mentioned type?

The answer must be in a Python list of length one. Examples of high-quality answers are: ['office chair'], ['dog'] or "['chair']". Please provide one noun phrase. Provide a short answer.

Figure 3: The textual prompt provided to ChatGPT agent to unify the responses produced by BLIP2 model and obtain a single class label per shape.

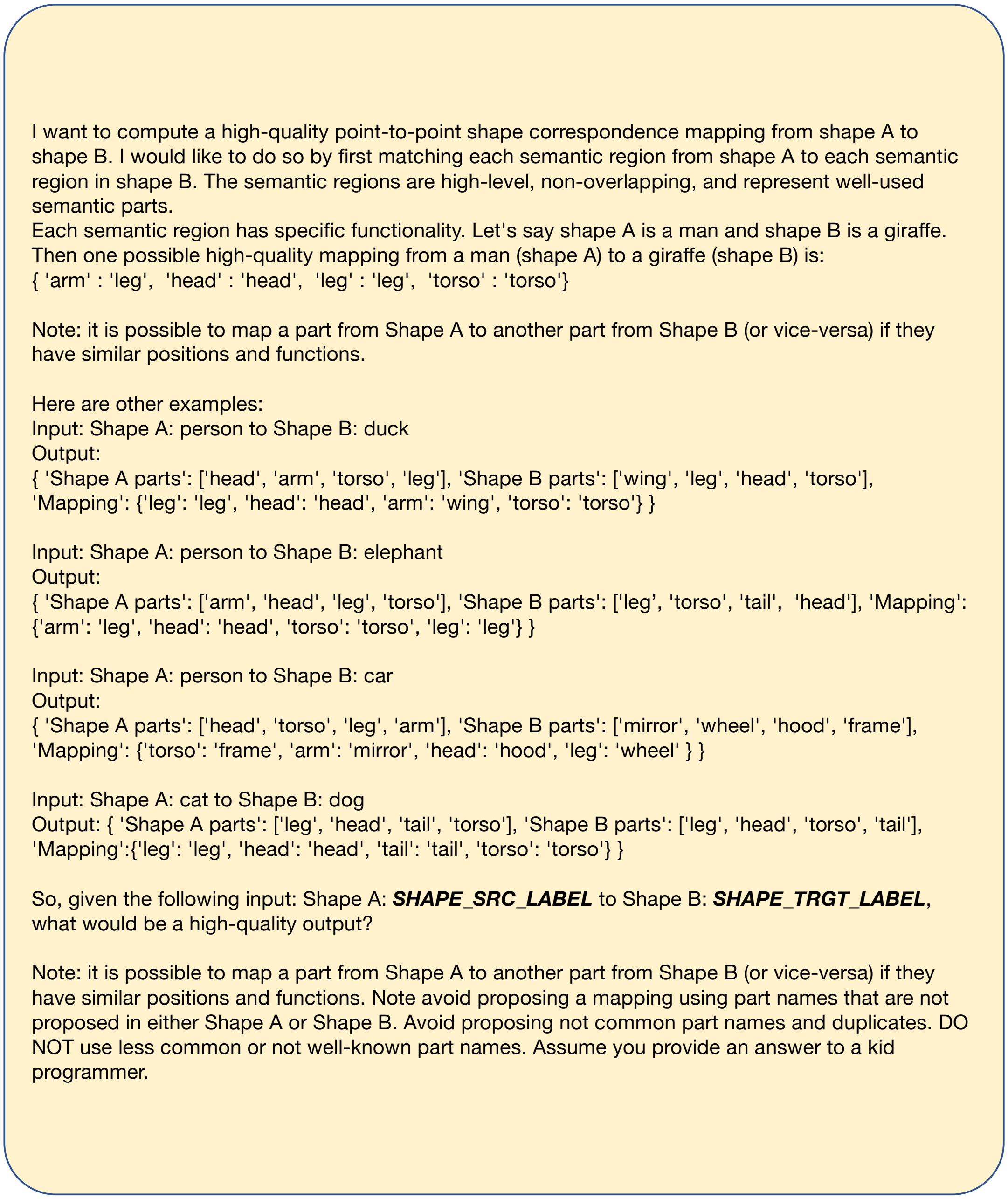
So, we construct a better prompt using only a single-step approach as described in Section 3.2, wherein the same prompt we ask ChatGPT to provide semantic regions for both input shapes and propose a mapping between the regions. In this manner, we can match region names that are different from each other but can be matched semantically.

D ZERO-SHOT 3D OBJECT CLASSIFICATION

We show in Figure 1 and Figure 3 the prompts used by BLIP2 and ChatGPT in our proposed method for zero-shot 3D object classification. In Figure 3, we replace the "ANSWERS_LIST" strings with a list of the proposals predicted by the BLIP2 model given the rendered images of an input shape.

D.1 GT Synonyms List

Figure 5 shows the collected synonyms we used in our proposed evaluation metrics.



I want to compute a high-quality point-to-point shape correspondence mapping from shape A to shape B. I would like to do so by first matching each semantic region from shape A to each semantic region in shape B. The semantic regions are high-level, non-overlapping, and represent well-used semantic parts.
Each semantic region has specific functionality. Let's say shape A is a man and shape B is a giraffe. Then one possible high-quality mapping from a man (shape A) to a giraffe (shape B) is:
{ 'arm' : 'leg', 'head' : 'head', 'leg' : 'leg', 'torso' : 'torso'}

Note: it is possible to map a part from Shape A to another part from Shape B (or vice-versa) if they have similar positions and functions.

Here are other examples:
Input: Shape A: person to Shape B: duck
Output:
{ 'Shape A parts': ['head', 'arm', 'torso', 'leg'], 'Shape B parts': ['wing', 'leg', 'head', 'torso'], 'Mapping': {'leg': 'leg', 'head': 'head', 'arm': 'wing', 'torso': 'torso'} }

Input: Shape A: person to Shape B: elephant
Output:
{ 'Shape A parts': ['arm', 'head', 'leg', 'torso'], 'Shape B parts': ['leg’, 'torso', 'tail', 'head'], 'Mapping': {'arm': 'leg', 'head': 'head', 'torso': 'torso', 'leg': 'leg'} }

Input: Shape A: person to Shape B: car
Output:
{ 'Shape A parts': ['head', 'torso', 'leg', 'arm'], 'Shape B parts': ['mirror', 'wheel', 'hood', 'frame'], 'Mapping': {'torso': 'frame', 'arm': 'mirror', 'head': 'hood', 'leg': 'wheel' } }

Input: Shape A: cat to Shape B: dog
Output: { 'Shape A parts': ['leg', 'head', 'tail', 'torso'], 'Shape B parts': ['leg', 'head', 'torso', 'tail'], 'Mapping':{'leg': 'leg', 'head': 'head', 'tail': 'tail', 'torso': 'torso'} }

So, given the following input: Shape A: ***SHAPE_SRC_LABEL*** to Shape B: ***SHAPE_TRGT_LABEL***, what would be a high-quality output?

Note: it is possible to map a part from Shape A to another part from Shape B (or vice-versa) if they have similar positions and functions. Note avoid proposing a mapping using part names that are not proposed in either Shape A or Shape B. Avoid proposing not common part names and duplicates. DO NOT use less common or not well-known part names. Assume you provide an answer to a kid programmer.

Figure 4: The textual prompt for proposing labels representing the semantic regions for an input pair of shapes and semantic region mapping using ChatGPT agent.

```
{
    "person": ["being", "body", "child", "creature", "human", "human being", "human body", "individual", "kid",
                "man", "soul", "woman"],
    "horse": ["colt", "cuddie", "cuddy", "dobbin", "filly", "gee-gee", "gelding", "hobby", "jade", "mare", "moke",
             "mount", "nag", "pony", "stallion", "steed", "stud", "studhorse", "yarraman", "yearling"],
    "fox": ["reynard"],
    "cougar": ["catamount", "mountain lion", "panther", "puma"],
    "lion": [],
    "wolf": [],
    "dog": ["brak", "canine", "cur", "hound", "kuri", "mongrel", "mutt", "pooch", "pup", "puppy", "tyke"],
    "cow": [],
    "hippo": ["hippopotamus"],
    "head": ["skull", "bean", "conk", "cranium", "crown", "loaf", "noddle", "noggin", "nut", "pate"],
    "arm": ["appendage", "upper limb"],
    "leg": ["limb", "lower limb", "member", "pin", "shank", "stump"],
    "tail": ["braid", "pigtail", "plait", "ponytail", "tress"],
    "torso": ["body"]
}
```

Figure 5: The collected synonyms for the ground-truth object classes and semantic regions we used in our proposed evaluation metrics.