

CusConcept: Customized Visual Concept Decomposition with Diffusion Models

1. Prompt Templates

To evaluate the performance of text-guided generation, we collect two sets of prompt templates for attribute concepts and object concepts respectively. Each set consists of 11 prompts:

Prompt templates for attributes:

- “a photo of S_* object”
- “a photo of S_* shirt”
- “a photo of S_* bed”
- “a photo of S_* leaf”
- “a photo of S_* clothes”
- “a photo of S_* street”
- “a photo of S_* carrot”
- “a photo of S_* bottle”
- “a photo of S_* house”
- “a photo of S_* car”
- “a photo of S_* woman”

Prompt templates for objects:

- “a photo of S_* ”
- “a photo of S_* at the beach”
- “a photo of S_* in the jungle”
- “a photo of S_* in the snow”
- “a photo of S_* in the street”
- “a photo of S_* on top of a pink fabric”
- “a photo of S_* on top of a wooden floor”
- “a photo of S_* with a city in the background”
- “a photo of S_* with a mountain in the background”

- “a photo of S_* with the Eiffel tower in the background”
- “a photo of S_* floating on top of water”

In our evaluation, we generate 8 images for each concept (attribute or object) and per prompt, resulting in 896 images in total. This extensive set allows for a comprehensive assessment of the method’s performance and its generalization capabilities.

2. Implementation details

Dataset. In the main paper, we collect a dataset for evaluating our model, comprising 56 images sourced from the VAW-CZSL dataset [4]. VAW-CZSL annotates each image with an object label and an attribute label describing the object. These annotations serve as ground-truth labels for our dataset, encompassing 14 attributes and 33 objects. The 14 attributes can be categorized into 8 attribute axes, with the attributes within each axis listed below:

- age: young
- busy: busy, calm
- cleanness: clean, dirty
- cook: fried, raw
- damage: damaged
- mood: laughing, sad
- straightness: straight, bent
- time: ancient, modern

The object labels are:

- boy, girl, lady, male, man, walkway, water, bed, carpet, fence, hair, keyboard, kitchen, door, face, glove, chicken, fries, broccoli, carrot, fridge, guy, person, cat, dog, phone, tree, pillar, pole, road, statue, suitcase, kitchen

LLM. For each attribute axis, we use GPT-4 [1] to generate 22 words. We do not need excessive words because the precise target word is not required in the first step, and fine-tuning in the second step can improve image accuracy.

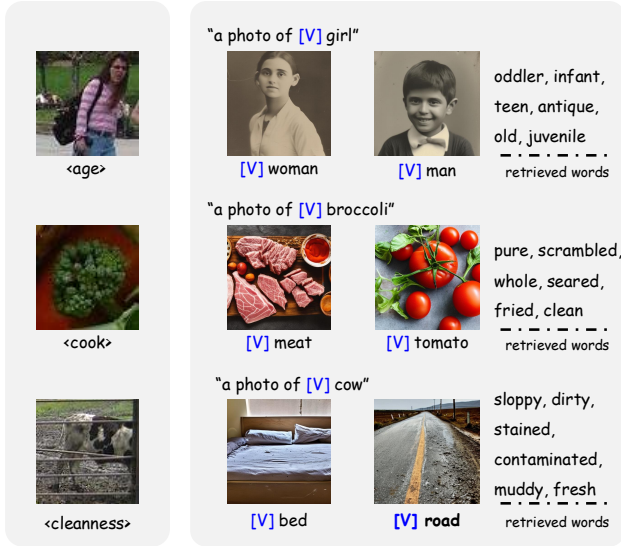


Figure 1. Generated images and retrieved words using our method on Dreambooth [3].

3. Deployment on Dreambooth

We have observed that similar to Textual Inversion [2], DreamBooth [3] also inserts token into the text to improve the fidelity and quality of the generated images. We applied our method to DreamBooth, and the results are shown in Fig. 1. Here, we replaced [V] in “a photo of [V] dog” (example in the DreamBooth diagram) with one of our attribute axes.

Both DreamBooth and Textual Inversion with our method can retrieve the correct words and generate images of the target attribute axes, effectively decomposing the original image. To maintain consistency with DreamBooth, we retain the original term ‘object’. However, this does not significantly improve the quality of the generated images, which likely due to the low pixel quality of the input images (from VAW-CZSL [4]).

4. Additional Visualized Results

CusConcept demonstrates remarkable abilities in disentangling and generating visual concepts. Here, we present additional visualization results that include the retrieved words from the vocabulary and generated images, featuring cases with five attribute axes in Fig. 2 and cases with three attribute axes in Fig. 3.

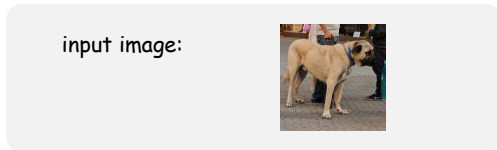
5. Societal Impact

This work aims to provide users with an effective tool for disentangling and generating visual concepts along user-specified attribute axes of human interest. While some previous works have focused on visual concept decomposi-

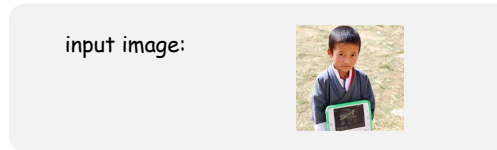
tion, we are the first to propose that users can specify arbitrary attribute axes, and the input image will be decomposed along these axes. Our method can also be used as a novel and effective way for concept removal and concept re-composition. However, our model may be maliciously employed for unethical purposes, such as generating violent, pornographic, or privacy-compromising content. We believe that legal compliance and open-source licenses are important for regulating the development of personalized generative models.

References

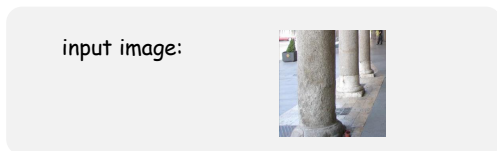
- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1
- [2] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2
- [3] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 2
- [4] Nirat Saini, Khoi Pham, and Abhinav Shrivastava. Disentangling visual embeddings for attributes and objects. In *CVPR*, 2022. 1, 2








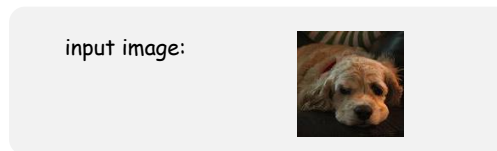
<p><color> a photo of S. bed</p> <p>ash, sandy, auburn</p> 	<p><hair> a photo of S. bed</p> <p>glossy, dull, dense</p> 
<p><wetness> a photo of S. clothes</p> <p>greasy, airy, oily</p> 	<p><mood> a photo of S. girl</p> <p>sad, sad, weeping</p> 
<p><cleanness> a photo of S. house</p> <p>sanitary, contaminated, spotted</p> 	<p>object a photo of S* on top of a wooden floor</p> <p>dog, beige, animal</p> 



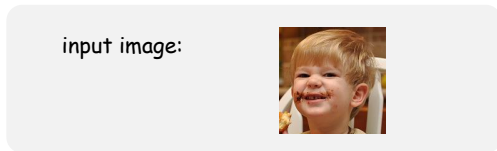
<p><color> a photo of S. clothes</p> <p>auburn, chestnut, mahogany</p> 	<p><haircut> a photo of S. woman</p> <p>bun, ponytail, long</p> 
<p><mood> a photo of S. woman</p> <p>lonely, sad, sad</p> 	<p><age> a photo of S. girl</p> <p>young, mature, senior</p> 
<p><cleanness> a photo of S. house</p> <p>stained, spotted, cloudy</p> 	<p>object a photo of S* with the Eiffel tower in the background</p> <p>korean, children, mongolian</p> 



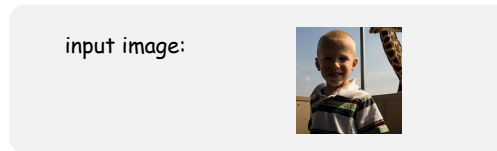
<p><color> a photo of S. bed</p> <p>indigo, gray, purple</p> 	<p><material> a photo of S. house</p> <p>wood, copper, stone</p> 
<p><cleanness> a photo of S. bed</p> <p>pure, sanitary, fresh</p> 	<p><straightness> a photo of S. bed</p> <p>upright, twisted, direct</p> 
<p><completeness> a photo of S. bottle</p> <p>perfect, solid, immaculate</p> 	<p>object a photo of S* on top of a wooden floor</p> <p>posts, pillars, statues</p> 



<p><color> a photo of S. bed</p> <p>ash, chestnut, sandy</p> 	<p><hair> a photo of S. bed</p> <p>silky, fluffy, dull</p> 
<p><wetness> a photo of S. car</p> <p>rainy, misty, dry</p> 	<p><mood> a photo of S. girl</p> <p>weeping, depressed, depressed</p> 
<p><cleanness> a photo of S. bed</p> <p>clean, cleaned, dusty</p> 	<p>object a photo of S* on top of a wooden floor</p> <p>rudolph, grover, hound</p> 



<p><color> a photo of S. bed</p> <p>ash, sandy, platinum</p> 	<p><haircut> a photo of S. girl</p> <p>ponytail, highlights, fringe</p> 
<p><mood> a photo of S. girl</p> <p>angry, amused, disappointed</p> 	<p><age> a photo of S. girl</p> <p>newborn, infant, baby</p> 
<p><cleanness> a photo of S. bed</p> <p>cloudy, messy, cleaned</p> 	<p>object a photo of S* with a mountain in the background</p> <p>children, child, kids</p> 



<p><color> a photo of S. girl</p> <p>blonde, silver, brown</p> 	<p><haircut> a photo of S. woman</p> <p>highlights, bald, bangs</p> 
<p><mood> a photo of S. girl</p> <p>sad, angry, amused</p> 	<p><age> a photo of S. girl</p> <p>teenage, adolescent, childish</p> 
<p><cleanness> a photo of S. shirt</p> <p>cloudy, sanitary, spotted</p> 	<p>object a photo of S* floating on top of water</p> <p>animal, kid, child</p> 

Figure 2. Visualization results along five attribute axes.

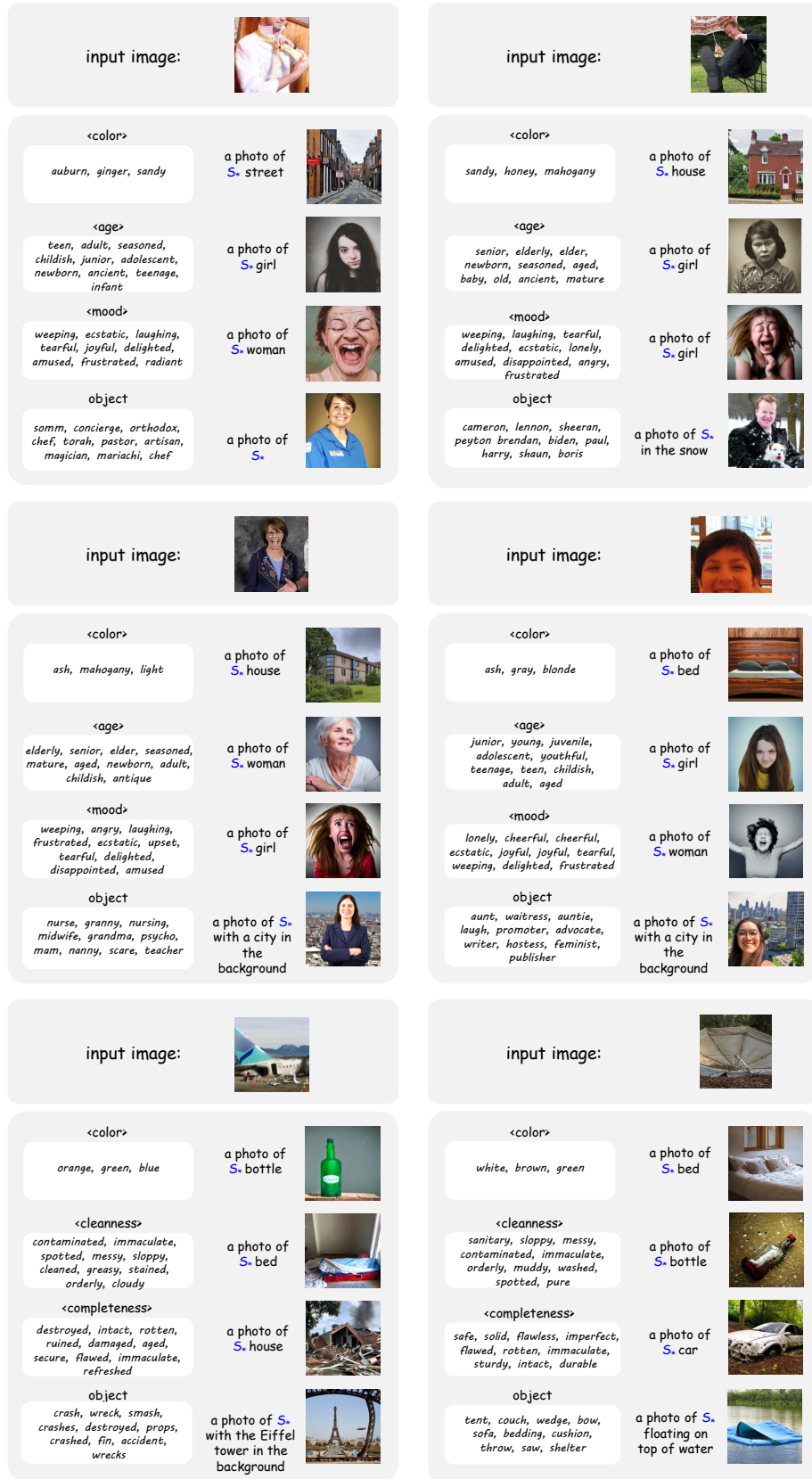


Figure 3. Visualization results along three attribute axes.