Appendix

In this appendix, we first show additional results of both decomposition and composition in appendix A. We then provide details of datasets used in our experiments in appendix B. Finally, we demonstrate details of baselines in appendix C and our method in appendix D, respectively.

A. Additional Results

In this section, we first provide analyses of the performance of our method on the sensitivity of the number of concepts K, the variance of our method on inferred concepts, and the diversity of generated images in appendix A.1. We then show additional results of decomposed concepts for objects, indoor scenes, artistic paintings, and hybrid dataset that consists of different modalities in appendix A.2. Finally, we provide additional results for object composition, indoor scene composition, art composition, and external composition in appendix A.3. Note that we utilize the conjunction operator (*e.g.*, AND) from composable diffusion [32] for compositional generation.

A.1. Analysis

Sensitivity of the number of concepts K. In Figure 11, we run our method with varying values of K (4, 5, and 6) on ImageNet S_1 (images with five categories of objects). When $K = 5$, our method can correctly find all five concepts. When $K < 5$, our method selects the top K obvious concepts. When $K > 5$, our method tries to discover some new concepts from the training data.

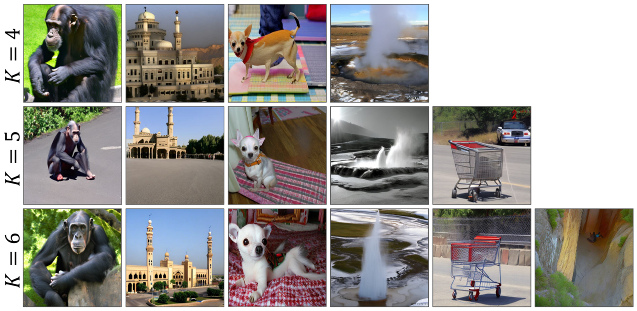


Figure 11: **Sensitivity of the number of concepts K on ImageNet Subset S_1.**

Variance of inferred concepts. We provide both qualitative and quantitive results on ImageNet S_1 across different seeds to assess the variance of inferred concepts. In Figure 12, we show that our method can reliably discover all object categories in S_1 across different seeds. In Table 3, we compare our method with the best baseline, *i.e.*, Textual Inversion (CKM) on ImageNet S_1. The result shows that our method can capture concepts consistently across multiple runs, as evidenced by higher accuracy, lower KL Divergence and smaller standard deviation values. The result is also consistent with the qualitative results in Figure 12.

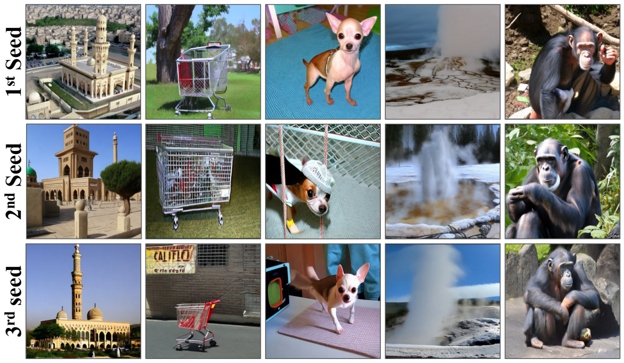


Figure 12: **Qualitative results on ImageNet S_1 across 3 random seeds.**

Models	ResNet-50		CLIP	
	Acc (%) ↑	**KL ↓**	**Acc (%) ↑**	**KL ↓**
TI (CKM)	36.35 ± 11.34	0.1432 ± 0.0637	34.16 ± 14.55	0.1386 ± 0.0101
Ours	$\mathbf{51.25 \pm 4.11}$	$\mathbf{0.0736 \pm 0.0626}$	$\mathbf{45.94 \pm 2.94}$	$\mathbf{0.0720 \pm 0.0969}$

Table 3: **Quantitative Evaluation on S_1 across 3 seeds.**

Diversity of generated images. In Table 4, we measure diversity in both training data and generated images by computing pairwise dot product similarity using CLIP, since CLIP is trained to differentiate similar and dissimilar data, thus allowing us to measure diversity or dissimilarity. Our result shows that greater similarity in training images leads to less diversity in the generated images.

A.2. Unsupervised Concept Discovery

Object Discovery. We show qualitative comparisons between our method and baselines for each set of ImageNet experiments. We find that baselines generate repetitive concepts. For example, textual inversion (KM) discovers two embeddings for the class of chimpanzee, as shown in the 3rd and 4th columns of the 1st row in Figure 15. Furthermore, both variants of textual inversion fail to generate certain concepts, such as *shopping cart*, while our method can discover such concepts, as shown in the 1st column of the Figure 15. We demonstrate that such problems exist across all experiments in Figure 15 and Figure 16. We also train COMET on ImageNet to decompose images into object categories. However, it scales poorly to more complex images, thus failing to decompose such images into realistic concepts as illustrated in Figure 17.

Indoor Scene Discovery. To further verify effectiveness of our approach, we provide additional qualitative results of our method on indoor scene decomposition, specifically in the kitchen setting. In Figure 18, we show both generated samples (odd columns) along their cross-attention maps (even columns) on three major concepts, incluing *kitchen range*, *kitchen islands*, and *lighting effects*.

Dataset	CLIP (Training Data)	CLIP (Generation)
ADE20K	0.1760	0.1701
Van Gogh	0.1411	0.1259
ImageNet S_1	0.1089	0.1188

Table 4: **Quantitative Evaluation on Image Diversity.**

Artistic Concept Discovery. Finally, we show our decomposed concepts based on artistic paintings, including Van Gogh (Figure 19), Claude Monet (Figure 20) and Pablo Picasso (Figure 21). Our method can discover artistic concepts from few paintings. We provide names of original paintings on the leftmost side for easy understanding.

Concept Discovery from hybrid modalities. We run our method on a hybrid dataset that contains images from four concepts, *i.e.*, kitchen, Geyser, Chihuahua, and Claude Monet paintings. As shown in Figure 13, our method can successfully discover all four distinct concepts.

A.3. Composing Discovered Concepts

Object Composition. We show our method enables multi-object composition in Figure 22. For example, we can generate images that resemble "a teddy bear sitting on a studio couch" in the 3rd row by composing two discovered classes.

Scene Composition. In Figure 23, we further compose indoor kitchen components to generate indoor scenes that contain given specifications, including combinations of *kitchen range* and *lighting effects* in 2nd row.

Style Composition. We also demonstrate compositioanl results of decomposed concepts from artistic paintings in Figure 24. In this experiment, we either compose concepts discovered from the same artistic (*e.g.*, 1st row) or even combine concepts across different artists (*e.g.*, 3rd row).

External Composition. Finally, we provide additional results of external composition, where we compose existing concepts (*e.g.*, text) with discovered concepts in Figure 25. We show that we can enable style transfer by composing text descriptions shown in the 1st column and discovered concept in the 2nd column to generate images.

Composition of multiple concepts. Our method can compose more than 2 concepts. In Figure 14, we show the composition of 3 concepts discovered from ImageNet S_1.

B. Details of Datasets

ImageNet [8]. We use 4 sets of ImageNet class combinations, denoted as ImageNet S_1, S_2, S_3 and S_4 in our experiments. Each combination consists of 5 object categories. S_1 includes *geyser*, *Chihuahua*, *chimpanzee*, *shopping cart* and *mosque*. S_2 includes *guinea pig*, *warplane*, *castle*, *llama* and *volcano*. S_3 includes *convertible*, *starfish*, *studio couch*, *african elephant* and *teddy*. S_4 includes *koala*, *ice bear*, *zebra*, *tiger* and *giant panda*. We randomly choose 5 images per category for each set as our training data.

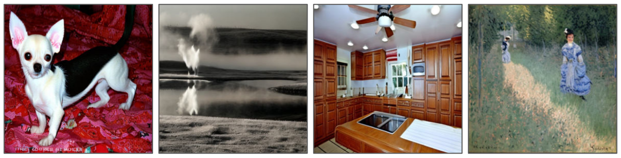


Figure 13: **Qualitative results on the hybrid dataset.**

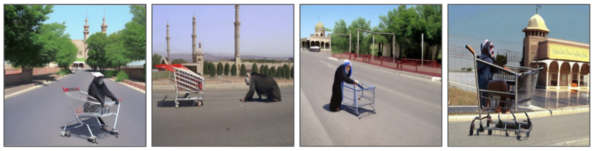


Figure 14: **Object Composition of 3 discovered concepts.**

ADE20K [56]. We use images in the *bedroom* subcatergory under the category of *home or hotel* for our training. Similarly, we randomly select 25 images as our training dataset.

C. Details of Baselines

COMET [11]. Most relevant to our work, COMET uses a set of EBMs to discover concepts in an unsupervised manner. However, COMET decomposes each individual image into a set of concepts, while our method decomposes a set of images into a set of concepts. Hence, COMET doesn't enable novel generation of the decomposed concepts and we instead visualize decomposed components from training images. For training, we use 5 components representing 5 object categories to train COMET using the default training setting from the official codebase.

Textual Inversion [15]. Given a set of similar images, textual inversion optimizes a single concept c, thus assuming a correspondence between the training data and the concept. In our experiments, however, we train an unconditional textual inversion by optimizing one single concept using all training images regardless of image classes or concepts. During inference, we generate 320 images using the prompt: "a photo of c" for evaluation.

Textual Inversion (KM). In this paper, our goal is to discover multiple concepts in an unsupervised way. Thus, we utilize unsupervised algorithms, such as K-means clustering, to obtain pseudo-labels. Before training a textual inversion model, we run K-means on the training images in pixel space to obtain predicted labels, which are used to optimize corresponding concepts during training. In our experiments, each ImageNet set has training images from 5 categories, so we initialize 5 concepts for optimization. During inference, we sample 64 images per concept for evaluation.

Textual Inversion (CKM). We also use another variant of textual inversion and K-means clustering as our baseline. In this case, we run K-means on image latent representations encoded by CLIP [38], thus we name it as CLIP-based K-means (*i.e.*, CKM for brevity). Similarly, we evaluate this baseline in the same way as textual inversion (KM).

Training Details. We train every single model with a batch size of 2 and 8 gradient accumulation steps, and thus an effective batch size of 16 per iteration for a total number of 3000 iterations on each dataset using a single NVIDIA A40 GPU. Other hyper-parameters (*e.g.*, optimizer) are the same as the original textual inversion codebase [15].

D. Details of Our Approach

Training. To discover compositional concepts we our approach, we initialize M (*i.e.*, 5) words along with their random embeddings as our concepts in our experiments, and a weight matrix with a shape of $N \times M$, where N is the number of training images. Then we utilize our method to optimize both weights and all M embeddings for each training image, as shown in Figure 2. Training details are the same as that of baselines shown in appendix C, where embeddings are optimized with a batch of 16 for 3000 iterations.

Inference. To enable image generation of each discovered concept, we sample images using each word using classifier-free guidance [21]. For compositional generation, we sample images using conjunction operator (*i.e.*, AND) from composable diffusion [32].

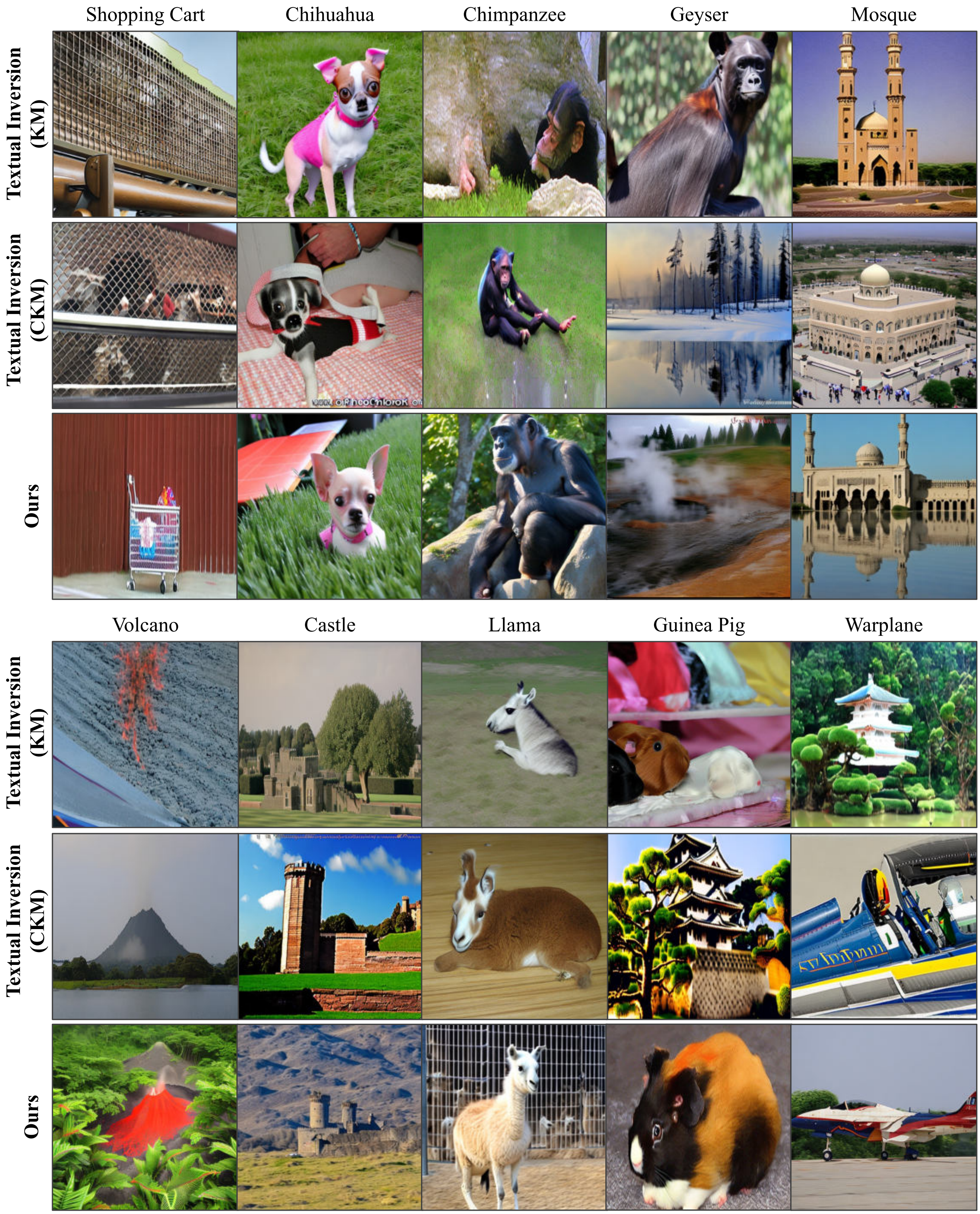


Figure 15: **Object Decomposition.** Object decomposition results on ImageNet S_1 (top) and S_2 (bottom). Note that concepts are labeled with our best interepretation for easy understanding.

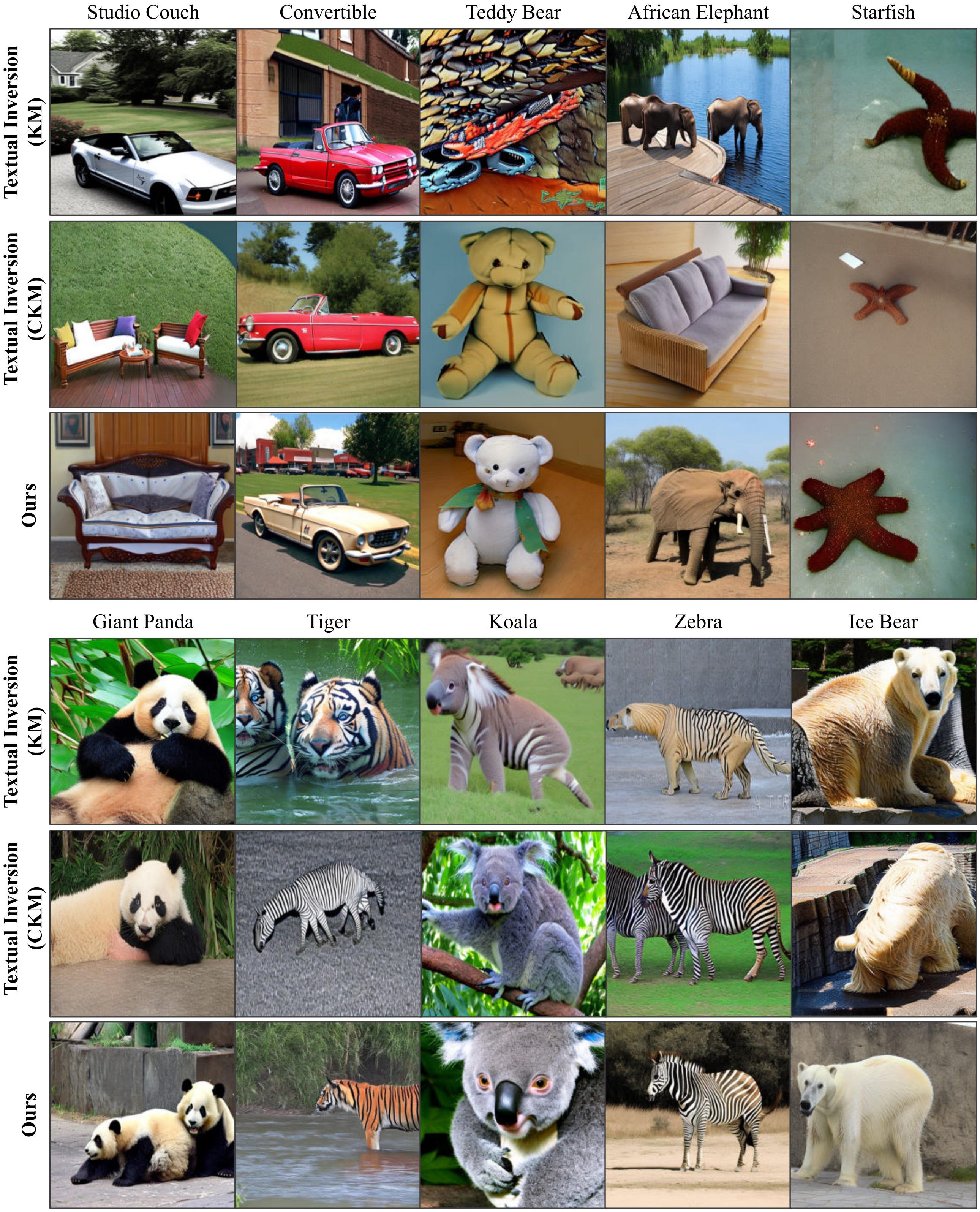


Figure 16: **Object Decomposition.** Object decomposition results on ImageNet S_3 (top) and S_4 (bottom). Note that concepts are labeled with our best interepretation for easy understanding.

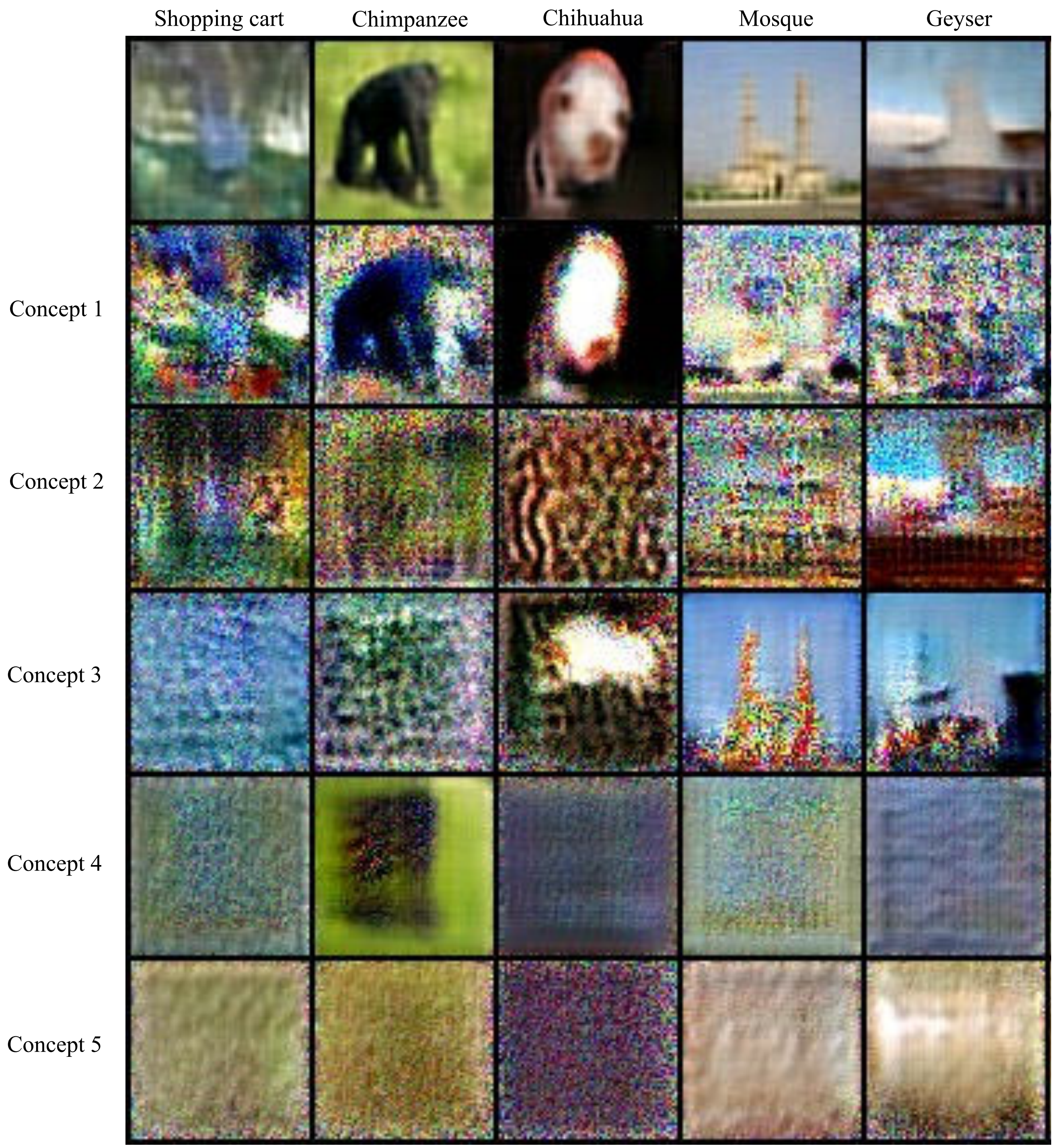


Figure 17: **Object Decomposition using COMET [11].** Object decomposition results on ImageNet S_1, where 5 of concepts learned from each training image (top) are not realistic.

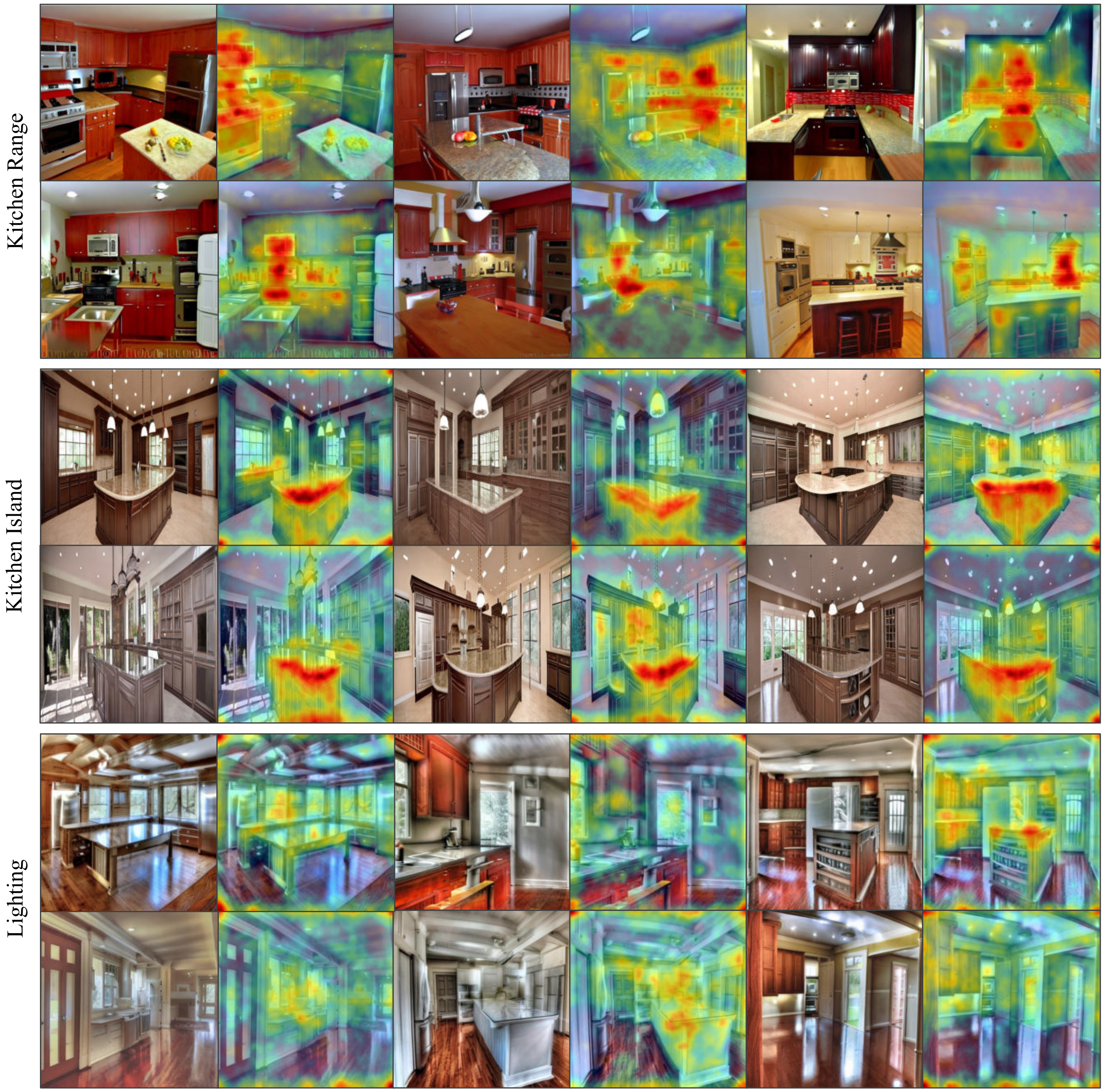


Figure 18: **Indoor Scene Decomposition.** We show additional results of decomposed kitchen concepts. Note that concepts are labeled with our best interepretation based on attention maps for easy understanding.

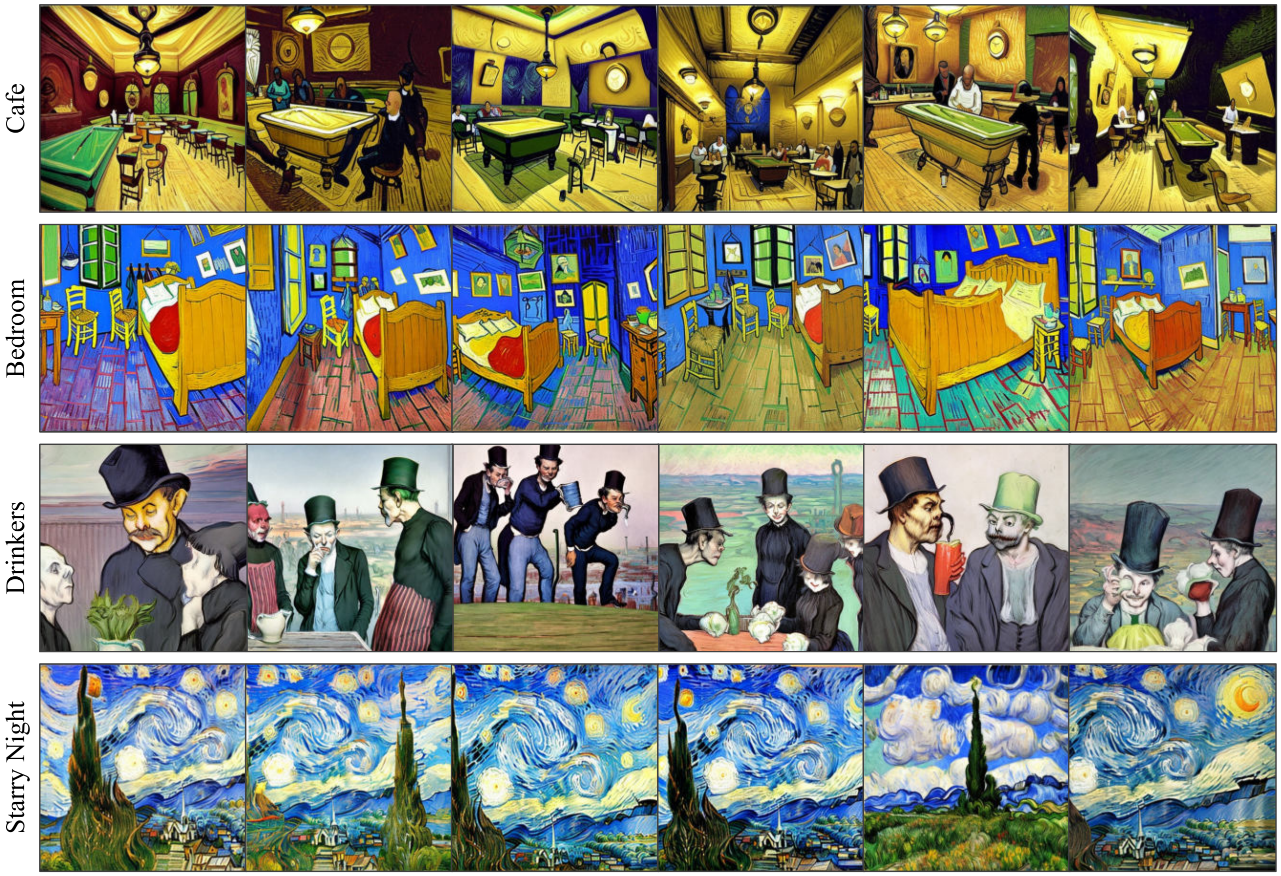


Figure 19: **Art Decomposition.** We show results of decomposed concepts using Van Gogh's paintings. Note that concepts are labeled with the name of the most similar paintings in the training set for easy understanding.

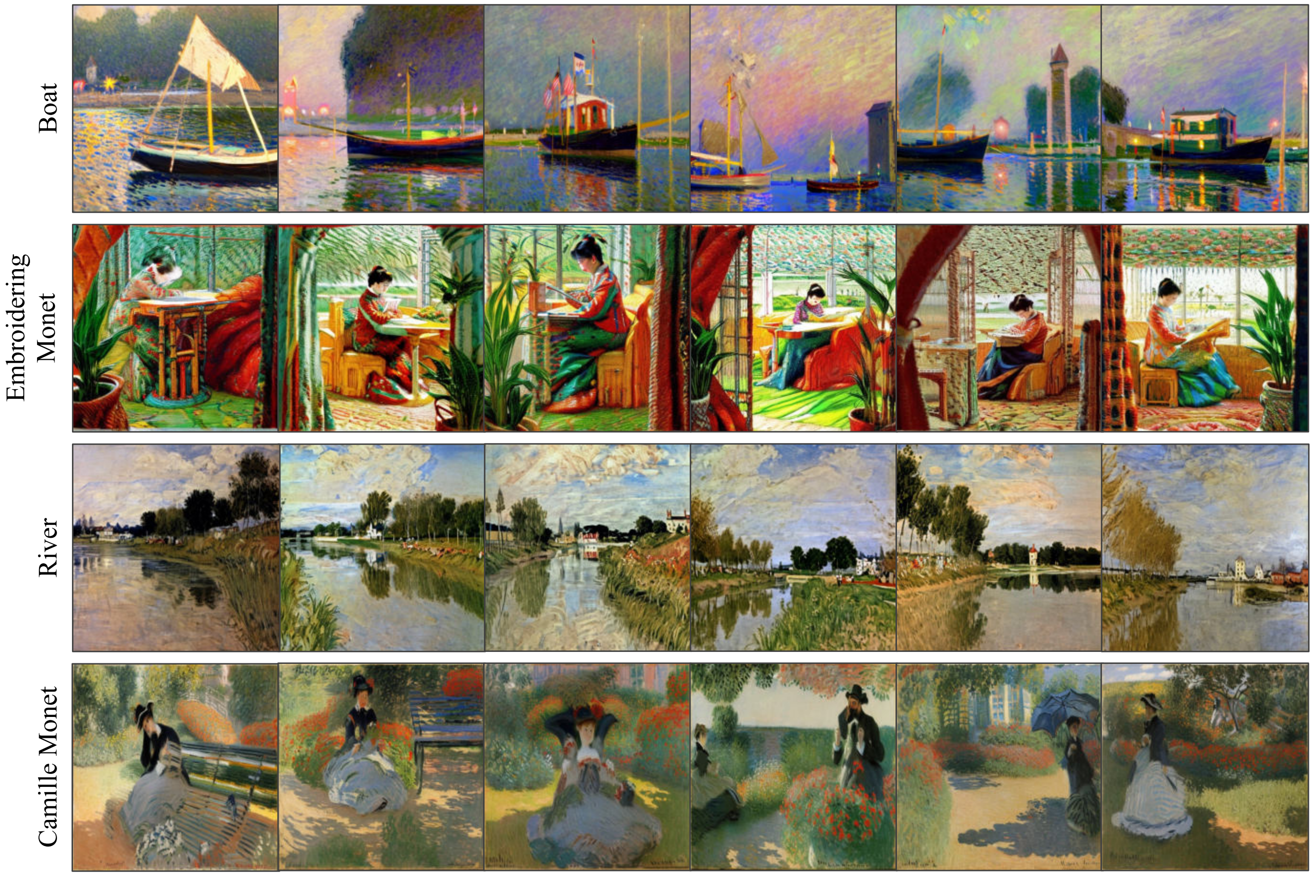


Figure 20: **Art Decomposition.** We show results of decomposed concepts using Claude Monet's paintings. Note that concepts are labeled with the name of the most similar paintings in the training set for easy understanding.

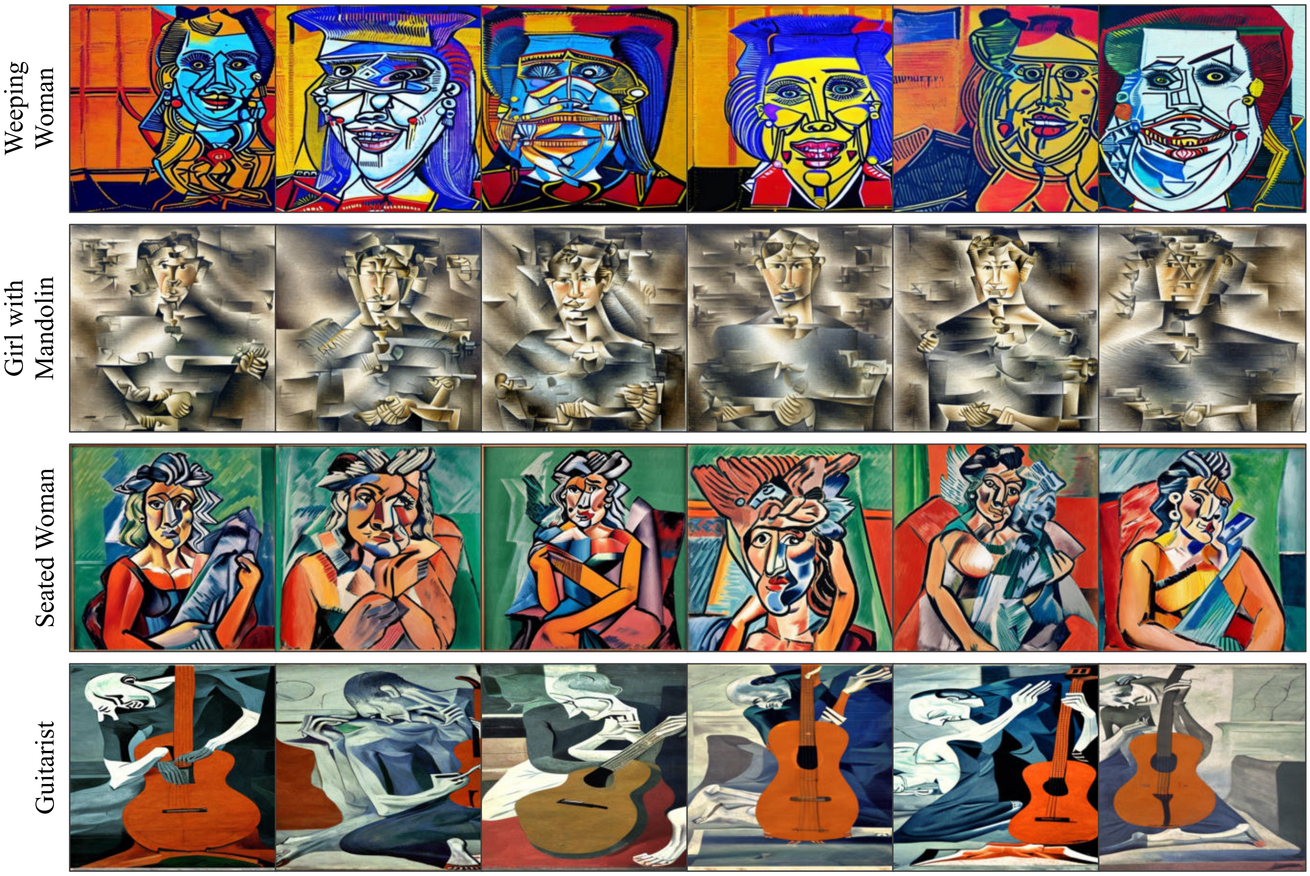
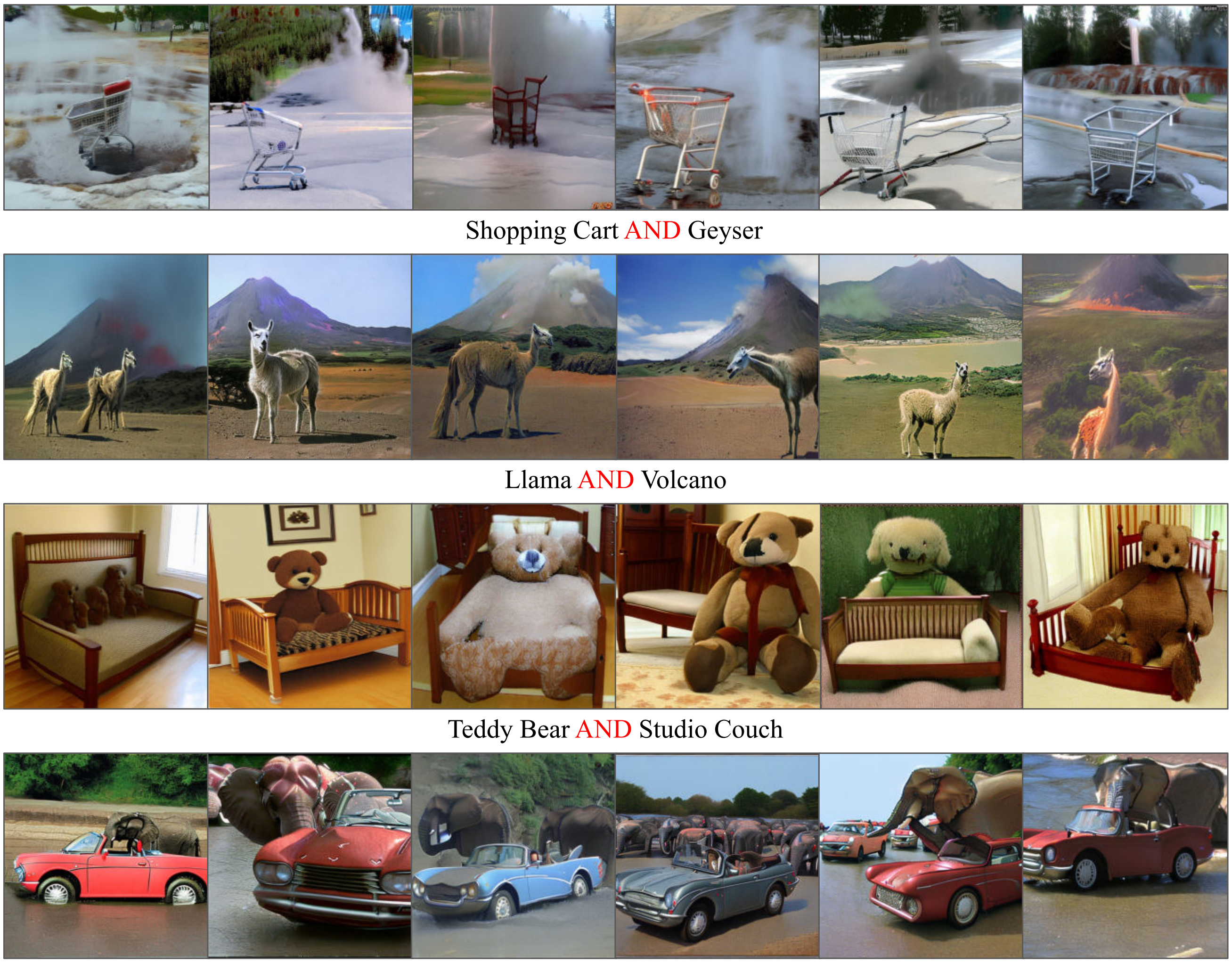


Figure 21: **Art Decomposition.** We show results of decomposed concepts using Pablo Picasso's paintings. Note that concepts are labeled with the name of the most similar paintings in the training set for easy understanding.



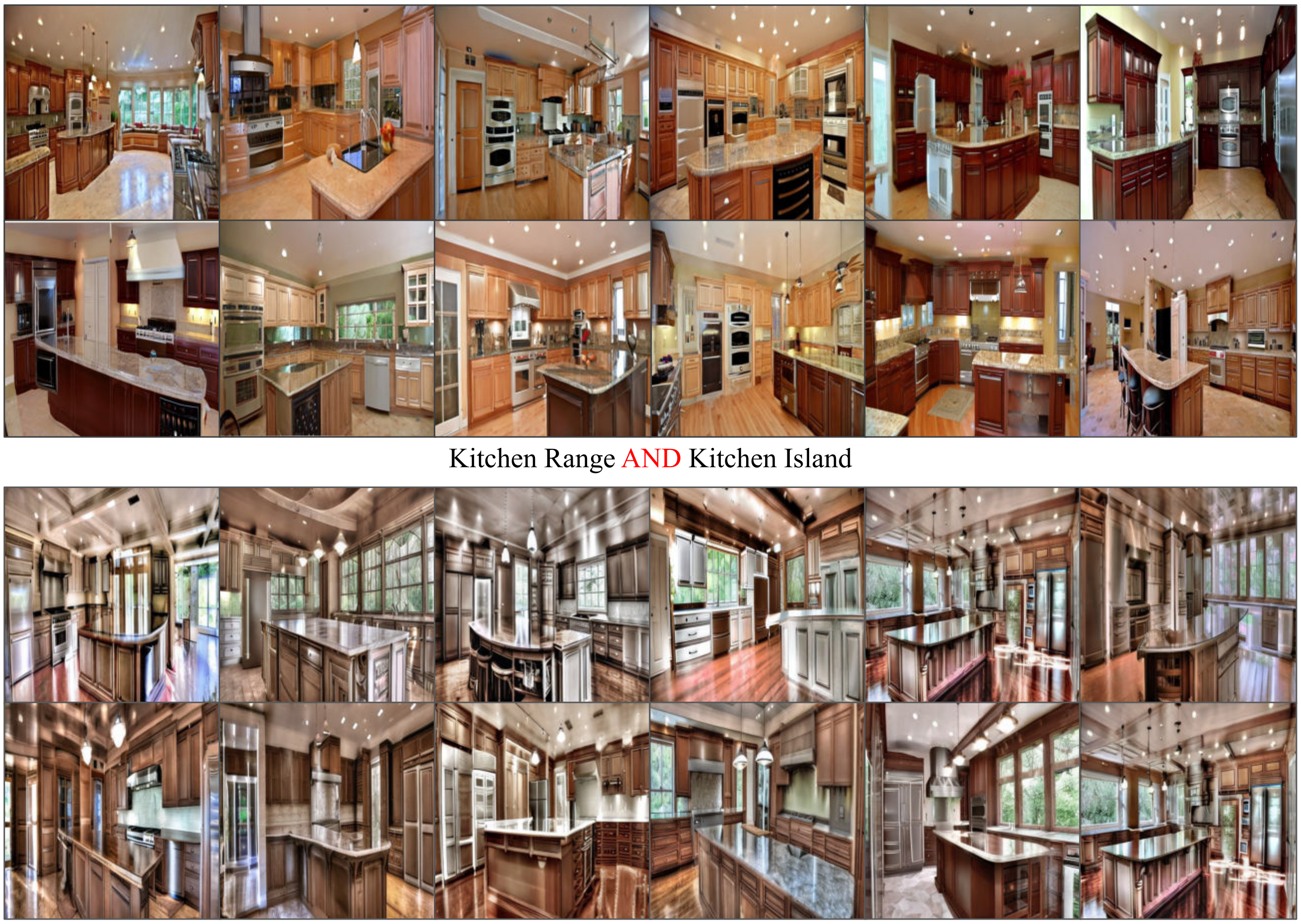
Shopping Cart AND Geyser

Llama AND Volcano

Teddy Bear AND Studio Couch

Convertible AND African Elephant

Figure 22: **Object Composition.** We show additional results of object composition using ImageNet classes. Note that concepts are labeled with our best interepretation for easy understanding.



Kitchen Range AND Kitchen Island

Kitchen Island AND Lighting

Figure 23: **Kitchen Scene Composition.** We demonstrate results of composing discovered kitchen components. Note that concepts are labeled with our best interpretation of what they are for easy understanding.

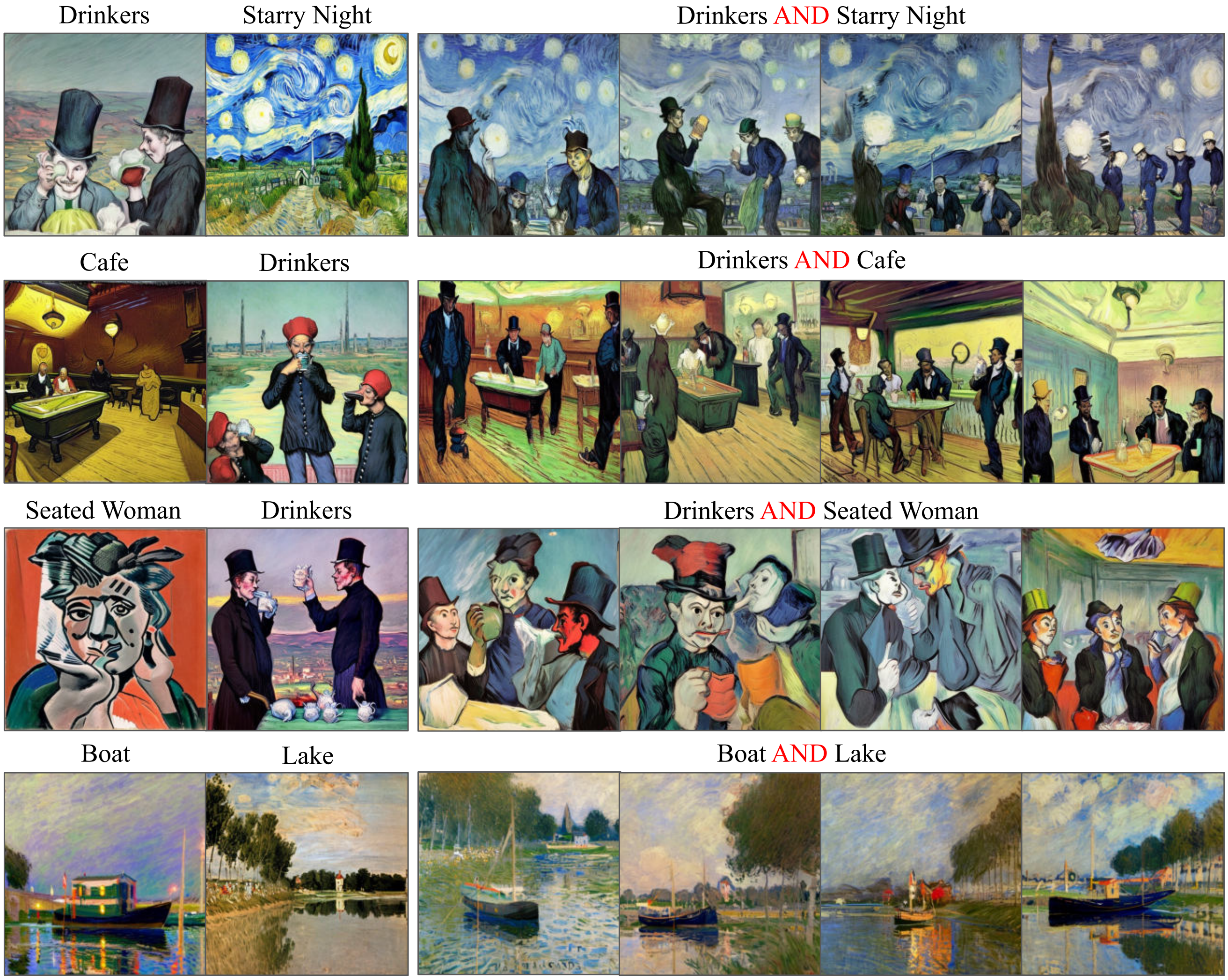


Figure 24: **Style Composition.** We show composition of different concepts discovered from differnet paintings. Note that concepts are labeled with the names of the most similar paintings in the training set.

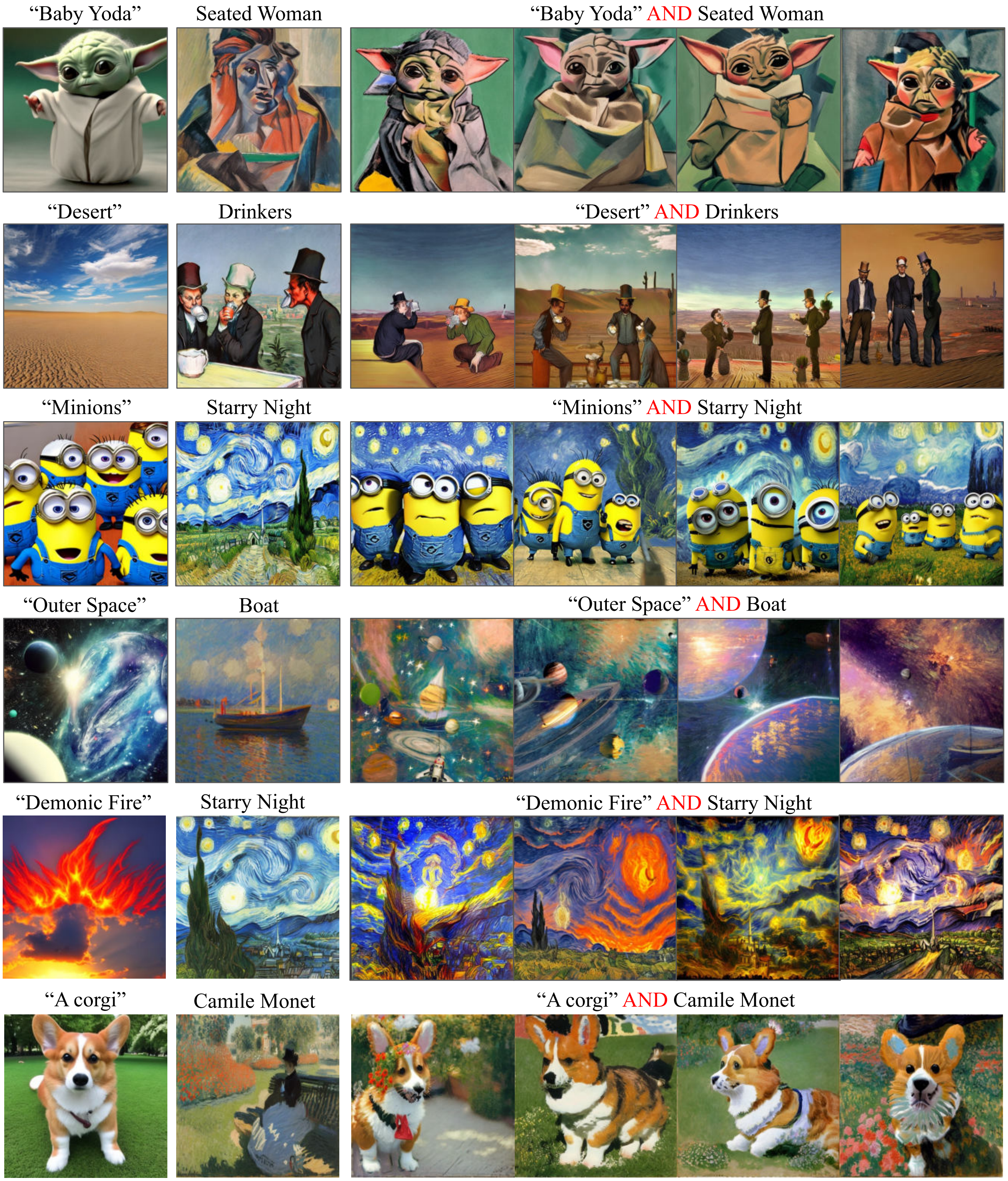


Figure 25: **External Composition.** We show composition results (3rd column) of existing concepts (1st column) and discovered concepts (2nd column), where discovered concepts are labeled with the names of the most similar paintings in the training set for easy understanding.