PartComposer: Composing Part-Level Concepts from Single-Image Examples

Junyu Liu R. Kenny Jones Daniel Ritchie
Brown University
{liu_junyu, russell_jones, daniel_ritchie}@brown.edu

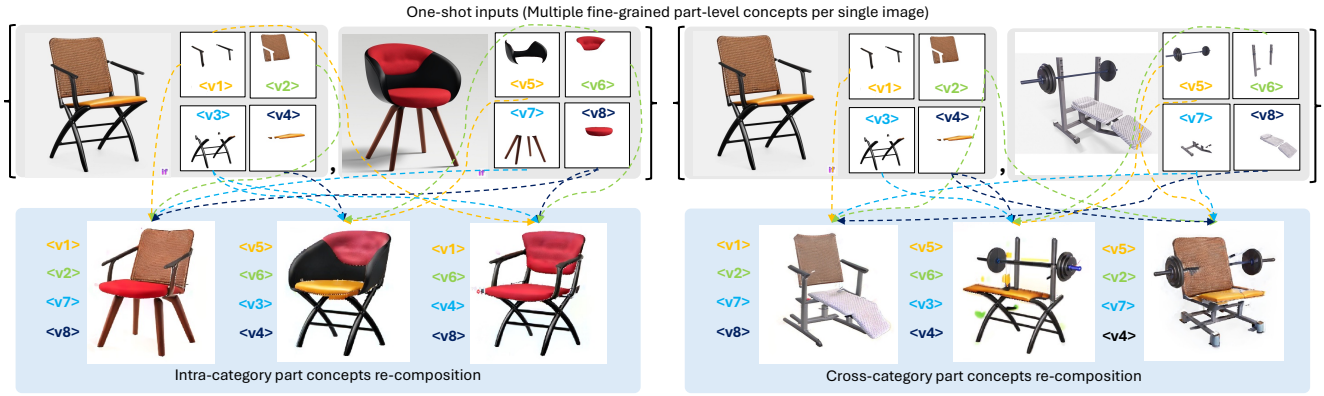

Figure 1. Given one-shot images that contains multiple fine-grained parts, our pipeline can learn descriptive concepts for these parts with disentanglement and flexibly re-compose them to generate new objects, for both intra-category and cross-category objects.

Abstract

We present a framework for part-level concept learning from single-image examples that enables text-to-image diffusion models to compose novel objects from meaningful components. Existing methods either struggle with effectively learning fine-grained concepts or require a large dataset as input. We propose a dynamic data synthesis pipeline generating diverse part compositions to address one-shot data scarcity. Most importantly, we propose to maximize the mutual information between denoised latents and structured concept codes via a concept predictor, enabling direct regulation on concept disentanglement and recomposition supervision. Our method achieves strong disentanglement and controllable composition, outperforming subject and part-level baselines when mixing concepts from the same, or different, object categories.

1. Introduction

Visually inspired and creative generation emerges from the ability to compose new objects from familiar parts [16, 17]. From virtual creatures to fantastical designs, part-level concept composition is a powerful paradigm for visual imagination. Recent works have explored extracting visual concepts from large generative models in the form of latent codes [5, 7, 10, 13, 19, 20, 22]. Each of the extracted concepts encodes the identity of an image object (e.g., a red seat cushion) and can be used with other concepts in creative image generation [2, 9, 11, 14, 16**?**]. In this context, text-to-image diffusion models serve as a versatile tool for learning compositional concepts through personalization. However, enabling these models to learn and compose **fine-grained part-level concepts** from just **single-image examples** (i.e., one image per object) remains challenging.

While recent methods advance concept learning with single-image input [10, 20, 22] or multi-concept capabilities [2, 6], they largely operate at the subject level and often fail to disentangle and retain the identity of fine-grained parts. Figure 2 illustrates this problem when composing 4 parts across 2 chairs. A common paradigm for these methods is to learn new or specialized token embeddings for concepts, through finetuning the diffusion model (updating the embeddings and/or the model weights) via latent diffusion loss $\mathcal{L}_{\text{ldm}}$. The learned concepts token embeddings can be directly used in prompts at inference time to generate images containing the target concepts. A common way to disentangle concepts is applying cross-attention loss $\mathcal{L}_{\text{attn}}$ between concept tokens, but this method alone cannot effectively deal with part-level concepts from one-shot inputs. Without regulating the information flow into these embeddings, part composition is prone to entanglement, ambiguity, or collapse. We present more detailed background discussion in Appendix A.

To overcome this issue, we propose a mutual information maximization framework that explicitly aligns denoised la-

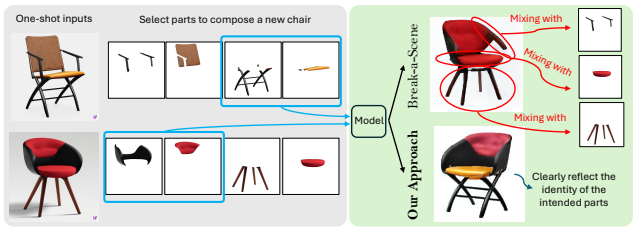

Figure 2. Illustration of the challenge in learning fine-grained concepts from one-shot inputs. Break-a-Scene [2] results in multiple entanglement while out method cleanly reflect target concepts.

tents with structured part-level concept codes. We introduce a concept predictor that performs both classification and segmentation on latent features, regularizing the embedding space to reflect concept presence and spatial structure. Additionally, we develop a dynamic data synthesis pipeline to generate rich part-level supervision from single-image examples.

Our framework enables compositional generation of new objects from single-image examples for the same and different object categories, advancing the capability of generative models to support creative, part-based visual reasoning.

2. Method

Given single-image examples, we want to be able to learn part concepts and mix them from each input together arbitrarily. Our pipeline build upon the commonly used diffusion model customization approaches (introduced in Section 1) like Break-a-Scene [2] and PartCraft [14], where standard diffusion loss $\mathcal{L}_{\text{ldm}}$ and cross-attention losses $\mathcal{L}_{\text{attn}}$ are used to learn and disentangle concept tokens. However, to enable learning concepts at part level with single-image examples, we propose a dynamic data synthesis method to augment the limited data and a maximizing mutual information scheme to enable clear disentanglement of parts' concepts and good composition capabilities. Figure 3 gives an overview of our approach.

Dynamic Data Synthesis We propose a dynamic data synthesis approach to augment the limited single-image example input. Our training batch contains two images, an instance image that is directly sampled from the given examples and a synthetic image that is generated on-the-fly. Figure 4 demonstrates our dynamic data synthesis approach. The instance image is randomly chosen from the input examples, and we randomly select a subset of the parts contained in the chosen instance image, inspired by the union sampling method in Break-a-Scene [2]. We mask out the unused parts areas and change the background to white to focus only on the part concepts, and pair it with a descriptive prompt. However, this data alone cannot provide enough training data to finetune diffusion models to mix different compositions of parts across different images. We thus propose to synthesize an image that contains parts that are randomly sampled across input examples. For each part category (e.g., the armrest of the chair), we randomly select a part from the input instances. Inspired by MuDI [9], we randomly scale and place the sampled parts on an white image. The parts can overlap each other to encourage learning concepts from multiple possible compositions. We modify the original masks according to the overlapping occlusions. Due to the high variability in the synthetic image, we use a different prompt to treat it as a collection of parts. Additional dynamic data synthesis explanation and examples can be found in Appendix D.

Maximizing Mutual Information Most existing concept learning methods either rely on multiple inputs [7, 13, 14] or are restricted by the number of concepts they could learn at the same time [10, 20, 22]. These restrictions results in poor performance in applying them on learning part-level concepts. We argue that a key missing point is that there are no explicit regulation on the information encoded in the concept embeddings. We thus propose a scheme to maximize the mutual information between the denoised latents and the concepts contained in the input image.

To this end, we adopt a mutual information maximization objective inspired by InfoGAN [3], which encourages the learned concept embeddings to retain interpretable and disentangled semantic structure. Let $\tilde{z}$ denote the denoised latent representation, and let $\mathbf{c}$ denote the set of concept codes associated with the input image. We seek to maximize the mutual information $I(\mathbf{c}; \tilde{z})$, which quantifies how much information about the concepts is preserved in the latent. This can be expressed as:

$$I(\mathbf{c}; \tilde{z}) = H(\mathbf{c}) - H(\mathbf{c} \mid \tilde{z}),$$

where $H(\mathbf{c})$ is the entropy of the concept distribution and $H(\mathbf{c} \mid \tilde{z})$ is the conditional entropy of the concepts given the latent. Since directly computing this term is intractable, we introduce a variational distribution $Q(\mathbf{c} \mid \tilde{z})$, implemented via a concept predictor, to approximate the true posterior. This leads to a variational lower bound:

$$\mathcal{I}_{\text{lower}} = \mathbb{E}_{\mathbf{c}\sim P(\mathbf{c}), \tilde{z}\sim P(\tilde{z}|\mathbf{c})}[\log Q(\mathbf{c} \mid \tilde{z})] + H(\mathbf{c}),$$

which satisfies $\mathcal{I}_{\text{lower}} \leq I(\mathbf{c}; \tilde{z})$. For a fixed prior over concepts (i.e., the ground-truth labels that supervise the concept predictor), $H(\mathbf{c})$ is constant and can be omitted during optimization. The corresponding training loss $\mathcal{L}_{\text{Info}}$ is defined as the negative of the lower bound which is minimized alongside the task loss. This encourages the model to produce latents $\tilde{z}$ from which the concept codes $\mathbf{c}$ can be accurately inferred, effectively regularizing the embedding space to reflect part-level semantics.

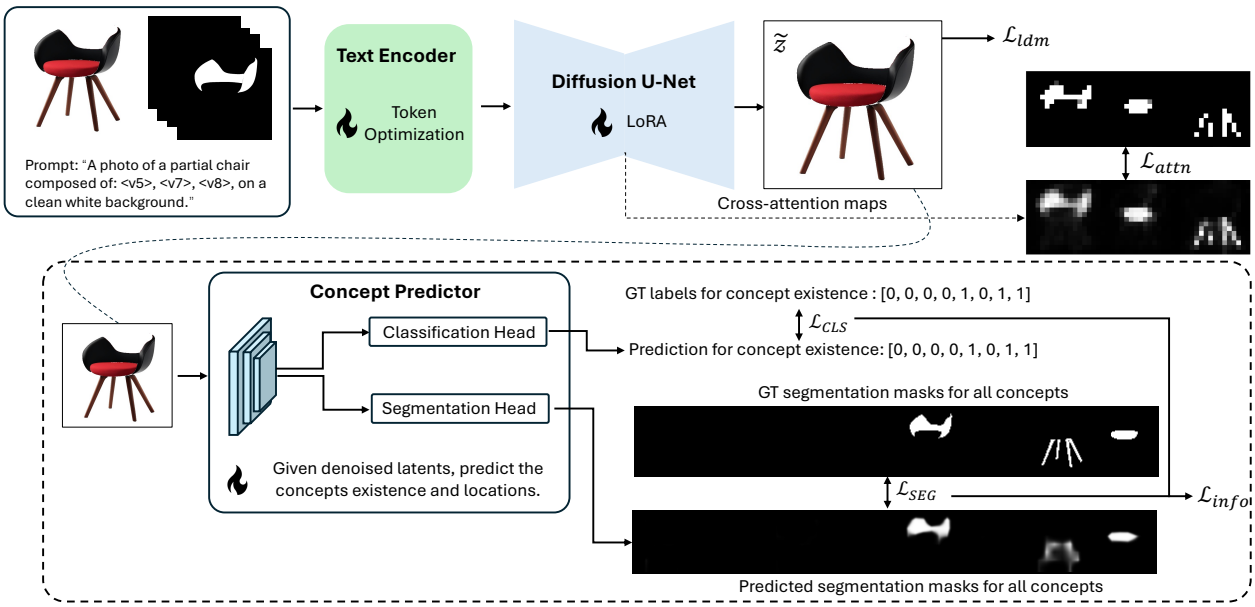

Figure 3. Overview of our method. Given a concept-compositional prompt and noise input, the denoising U-Net produces a latent $\tilde{z}$ supervised by reconstruction loss $\mathcal{L}_{\text{ldm}}$, cross-attention losses $\mathcal{L}_{\text{attn}}$, and information loss $\mathcal{L}_{\text{info}}$. $\mathcal{L}_{\text{info}}$ is computed by a concept predictor which receives $\tilde{z}$ and outputs concept classification and segmentation logits. The goal is to maximize mutual information between latent features and concept codes. All modules are jointly optimized to enable part-level concept disentanglement and controllable composition.

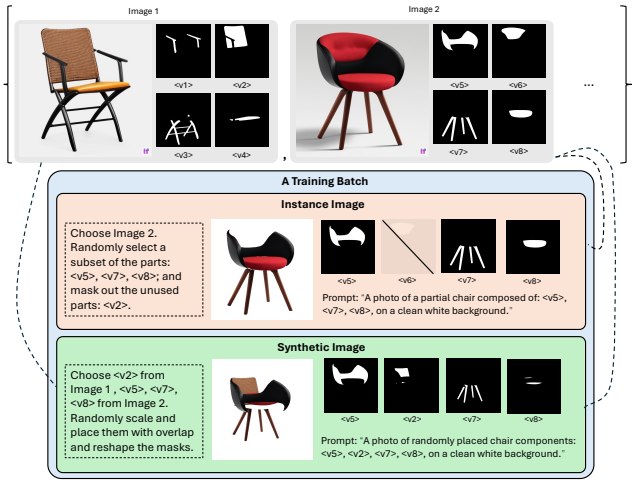

Figure 4. Dynamic data synthesis. Each training batch includes an instance image with masked parts and a synthetic image with randomly sampled and placed parts from multiple inputs, enabling diverse part-level supervision from single-image examples.

In our concept predictor design shown in Figure 3, we use two output heads to provide both classification and segmentation predictions of concepts given an denoised latent. The classification loss $\mathcal{L}_{\text{CLS}}$ penalizes the wrong compositions of concepts (i.e., missing concepts or containing more concepts). The segmentation loss $\mathcal{L}_{\text{SEG}}$ further penalizes wrong localization of concepts. These two losses are weighted to have the same scale and combined to get the mutual information loss $\mathcal{L}_{\text{Info}}$. The concept predictor is jointly optimized with the concept learning process. More detailed loss function descriptions and weight settings are introduced in Appendix D.

3. Experimental Results

The implementation details and experimental settings are presented in Appendix B. We present the qualitative results for part-level concept learning and composition (for intra- and cross-category objects) in this section.

Intra-category Results We first demonstrate the part-level concept learning and composition capability of our model using inputs from same categories. The input subjects consist same part decompositions (e.g., we decompose a chair into 4 parts: armrest, seat back, legs, and seat in Figure 5.) We compare the our methods with Break-a-Scene [2], which is representative for concept learning from a single image, and PartCraft [14], which is representative for part-level concept learning. To align their setting with our task, we adapt the original Break-a-Scene input dataset into a multi-image input manner and feed the single-image examples into PartCraft pipeline. More implementation details for the adaptation of their methods are explained in Appendix E.

Figure 5 shows the qualitative comparison of our method

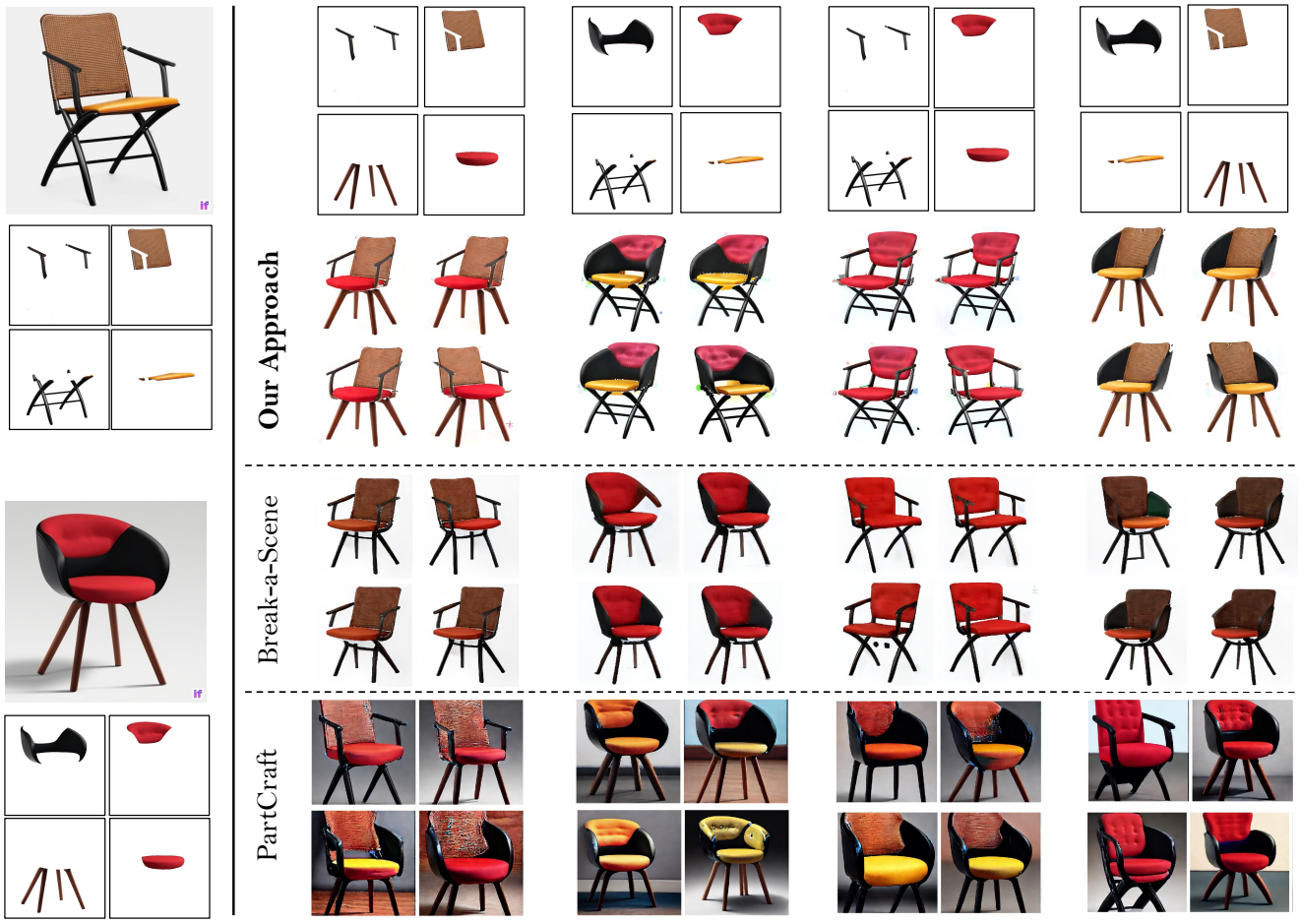

Figure 5. Comparison of Concept mixing results for 2 chairs using our approach, Break-a-Scene [2], and PartCraft [14]. The input images and correspond part decompositions are shown in the left column. We illustrates 4 part composition across the two input images by selecting 2 parts from each image, which are shown on the top row. We randomly sample 4 generated images for each composition.

with Break-a-Scene [2] and PartCraft [14], where we aim to recompose concepts across two input images. We generate 4 samples for each part composition. Break-a-Scene [2] struggles to disentangle and recompose part-level concepts, resulting in mixed identity in different parts, presumably since it was designed to do subject-level concept learning. PartCraft struggles to learn and recompose part-level concepts from single-image examples, resulting in poor and inconsistent image generation quality, presumably since it was designed to train on a large dataset. Our pipeline produce clear part-level concept disentanglement and clean composition capability. Additional intra-category results are presented in Appendix F.1.

Cross-category Results To enable more creativity in part-level concepts composition, we evaluate our methods in learning and remixing parts from objects in different categories, aiming to generate virtual objects. Figure 1 shows hybrid compositions from a chair and a gym equipment. More results are shown in Figure 17 in Appendix F.2, where the parts from a chair and a bed are hybridly composed. Our methods can generate creative virtual objects with different part compositions, preserving clear part-level identity with reasonable structural arrangements for most composition scenarios.

4. Conclusion

We propose a mutual information maximization frameworks in text-to-image diffusion model customization to disentangle the concepts and reduce concept missing in composing new objects. Together with our dynamic data synthesis approach, we enable learning and composing fine-grained part-level concepts from single-image examples. Our methods can produce identity-preserved and natural part compositions given the single-image examples from the same or different object categories.

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. 6

[2] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023. 1, 2, 3, 4, 6, 7

[3] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016. 2

[4] DeepFloydAI. Deepfloyd if, 2023. 6

[5] 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. 1, 6

[6] Daniel Garibi, Shahar Yadin, Roni Paiss, Omer Tov, Shiran Zada, Ariel Ephrat, Tomer Michaeli, Inbar Mosseri, and Tali Dekel. Tokenverse: Versatile multi-concept personalization in token modulation space. *arXiv preprint arXiv:2501.12224*, 2025. 1, 6

[7] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7323–7334, 2023. 1, 2, 6

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6, 7

[9] Sangwon Jang, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. Identity decoupling for multi-subject personalization of text-to-image models. *arXiv preprint arXiv:2404.04243*, 2024. 1, 2, 6

[10] Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*, 2023. 1, 2, 6

[11] Chen Jin, Ryutaro Tanno, Amrutha Saseendran, Tom Diethe, and Philip Alexander Teare. An image is worth multiple words: Discovering object level concepts using multi-concept prompt learning. In *Forty-first International Conference on Machine Learning*, 2024. 1

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 6

[13] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023. 1, 2, 6

[14] Kam Woh Ng, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Partcraft: Crafting creative objects by parts. In *European Conference on Computer Vision*, pages 420–437. Springer, 2024. 1, 2, 3, 4, 6, 7

[15] RenderHub. Renderhub. https://www.renderhub.com. 6

[16] Elad Richardson, Kfir Goldberg, Yuval Alaluf, and Daniel Cohen-Or. Conceptlab: Creative concept generation using vlm-guided diffusion prior constraints. *ACM Transactions on Graphics*, 43(3):1–14, 2024. 1

[17] Elad Richardson, Kfir Goldberg, Yuval Alaluf, and Daniel Cohen-Or. Piece it together: Part-based concepting with ip-priors. *arXiv preprint arXiv:2503.10365*, 2025. 1, 6

[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. https://github.com/Stability-AI/stablediffusion. 6, 7

[19] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 1, 6

[20] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8543–8552, 2024. 1, 2, 6

[21] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 7

[22] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953, 2023. 1, 2, 6

[23] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 6

A. Background in Visual Concept Learning

Input Requirements Pioneering works like Textural Inversion [5] and DreamBooth [19] in visual concept learning require multiple images as input to encode a single a concept. Several follow-up works have loosen the input constraint as learning single concept from single-image examples [10, 20, 22]. Multi-concept learning from multiple inputs are also being discovered [7, 13]. Break-a-Scene [2] first propose a pipeline to learn multiple concepts from single-image examples, enabling great flexibility in learning and remixing concepts from general use cases.

Concept Granularity Most concept learning methods focus on learning concept for the entire image or on the subject level [2, 6, 7, 9, 10, 13, 20, 22]. The target use case for these methods is to generate images that are organically composed by required subject(s) according to user-specified prompts. However, to enable more creativity in generating new subjects, part-level concept learning is required. This more fined-grained concept learning task has been explored by PartCraft [14]. They learn a large dictionary of concepts for different parts of creatures and generate new virtual creatures that have not been seen in the real world. However, their method rely on training on a large dataset and struggle on single-image examples. Piece-it-Together (PiT) [17] also targets part-level concept learning, by directly operating in a carefully chosen IP-Adapter+ [23] representation space and synthesizes a complete and coherent concept by training a generative model to fill in missing information conditioned on a strong domain-specific prior. Unlike optimization-heavy approaches, PiT enables efficient inference, supports diverse sampling from sparse inputs, and allows flexible semantic manipulation. However, PiT still requires training on class-specific datasets, where direct single-image examples are not supported.

Common Challenges A common challenge in multi-concept learning is to disentangle the identity of concepts [2, 9, 14]. To achieve good disentanglement, dynamic masking different compositions of concepts and using cross-attention mechanisms to regulate the diffusion model finetuning are commonly used in methods like Break-a-Scene [2] and PartCraft [14]. MuDI [9] proposes a dynamic concept composition method and mean-shifted inference technique to further improve the decoupling of different concepts. However, concepts missing or inaccurate identity still occur in all of these methods when remixing/recomposing concepts in generative process [2, 9, 14], especially when dealing with 4 or more concepts. Our observation is that all existing concept learning works do not provide any explicit regulation on the information encoded in different concepts, causing inaccurate multi-concept remixing. Other common challenges include learning large number of concepts (e.g., more than 4) from single inputs [2].

B. Experiments Setup

Data We use both synthetic data and real-world images in our experiments. We generate the synthetic images using DeepFloyd IF [4] and we collect real-world images mainly from renderings of 3D objects [15]. The mask generation methods are introduced in Appendix C.

Implementation Details For all experiments, we use Stable Diffusion v2.1 [18] as the pre-trained text-to-image diffusion model and apply LoRA [8] with rank 32 to the U-Net module. The concept predictor is a convolutional network that operates on denoised latents and outputs both classification and segmentation predictions. It consists of three convolution layers (with output channels 16, 32, and 64), followed by two parallel output heads: a classification head composed of two fully connected layers for predicting multi-label concept presence, and a segmentation head consisting of a 1×1 convolution followed by bilinear upsampling to produce per-concept spatial masks. We follow a two-stage training scheme similar to Break-a-Scene [2], where we only update the token embeddings with a high learning rate (10^{-4}) in the first stage and fully update both text encoder and LoRA weights in the U-Net with a low learning rate (10^{-6}) in the second stage. More training and inference details are introduced in Appendix D.

C. Mask Generation

Our method assumes access to part-level masks for training, but remains agnostic to how these masks are obtained. In practice, we adopt one of the following strategies depending on the dataset:

- **Automatic segmentation and labeling:** We apply off-the-shelf segmentation models such as SAM [12] to produce over-segmented masks, followed by GPT-4o-based [1] captioning and labeling to group and assign part identities.
- **Manual or direct annotation:** In some settings, we directly specify or provide part masks, either manually or from existing annotated assets (e.g., 3D renderings [15] or synthetic data).

Our framework is compatible with any source of part-level supervision, including both automatic and manual pipelines. Since the main focus of our work is not on mask generation itself, we treat this step as a pre-processing module and do not over-emphasis our effort on it.

D. Training Details

We provide the detailed training configuration used in our experiments for learning part-level visual concepts from

single-image examples. Our pipeline is trained using Stable Diffusion v2.1 [18] as the base text-to-image model. We use a pair of `chair` images from Figure 4 as an illustrative example. Each chair is decomposed into four semantic parts: `armrest`, `backrest`, `legs`, and `seat`. These parts are tokenized into 8 learnable placeholder tokens, denoted as `<v1>` through `<v8>`, with each token corresponding to a specific part instance from the two training images.

Prompt Design. Each training image is paired with a structured, part-compositional prompt using the assigned placeholder tokens. Two types of prompts are generated dynamically during training:

- **Instance prompts:** These describe partial objects composed of a subset of parts from a single input image. For example, if parts `<v5>`, `<v7>`, and `<v8>` are selected from Image 2, the corresponding prompt is: *"A photo of a partial chair composed of: <v5>, <v7>, <v8>, on a clean white background."*
- **Synthetic prompts:** These describe compositions of parts sampled across both input images. For example, if `<v2>` is sampled from Image 1 and `<v5>`, `<v7>`, and `<v8>` are sampled from Image 2, the prompt is: *"A photo of randomly placed chair components: <v2>, <v5>, <v7>, <v8>, on a clean white background."*

All prompts are automatically generated based on the selected part indices. Backgrounds are set to white by default, and prompts are templated consistently to ensure clean compositional control.

Loss Configuration. To enforce both generative quality and part-level semantic disentanglement, we optimize a weighted combination of losses:

- $\mathcal{L}_{\text{ldm}}$: The standard latent diffusion reconstruction loss.
- $\mathcal{L}_{\text{attn}}$: Cross-attention loss to promote concept disentanglement.
- $\mathcal{L}_{\text{CLS}}$: Multi-label classification loss from the concept predictor's classification head.
- $\mathcal{L}_{\text{SEG}}$: Per-pixel segmentation loss from the predictor's segmentation head.
- $\mathcal{L}_{\text{BG}}$: Background supervision loss that penalizes generation of content outside the union of selected part masks. This encourages the model to focus on concept-relevant regions and avoid unrelated artifacts in unmasked areas.

The total loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ldm}} + \mathcal{L}_{\text{attn}} + \lambda_{\text{cls}}\mathcal{L}_{\text{CLS}} + \lambda_{\text{seg}}\mathcal{L}_{\text{SEG}} + \lambda_{\text{bg}}\mathcal{L}_{\text{BG}},$$

with the following loss weights:

$$\lambda_{\text{cls}} = 0.05, \quad \lambda_{\text{seg}} = 10.0, \quad \lambda_{\text{bg}} = 0.01.$$

Training Procedure. We adopt a two-phase training scheme:

- **Phase 1:** Only concept token embeddings are optimized for 6,400 steps.
- **Phase 2:** The LoRA-injected U-Net (with rank 32) and text encoder are jointly fine-tuned for 40,000 steps.

We want to note that the second stage training step number is just a rough reference which generally ensures that all concepts are well-learned, and removes background contents. In general, after 18,000 steps, the concept learning and re-composition results are already good enough.

Inference Procedure. We perform inference using standard DDIM [21] sampling with a pretrained Stable Diffusion v2.1 backbone, the optimized text encoder, and the learned LoRA weights. Given a compositional prompt (e.g., `"A photo of a partial chair with <v2>, <v5>, <v7>, <v8>, on a clean white background."`), the model decodes a final image using 50 DDIM steps. We use a commonly used guidance scale 7.5.

Learning Extra Background Concepts. Our pipeline also supports learning extract background concepts if user wants to place the re-composed object on specific background images. We use `<bgX>` to represent the background concepts and they are learned together with the part-level concepts. In our data dynamic data synthesis stage, we replace the white background with the given background images and keep all other operations as the same. We use the background loss λ_{bg} with weight $\lambda_{\text{bg}} = 0.01$ to train the background concepts. Figure 6 show the results for incorporating an indoor background with re-composed chairs. Our method can naturally blend the newly composed chair into the given background.

E. Qualitative Comparison Implementation

We compare our method with other visual concept learning and re-composition works, Break-a-Scene [2] and PartCraft [14]. We introduce the detailed comparison experiment setup in this section.

Break-a-Scene [2] is originally designed to learn mutiple subject-level concepts in a single image. We modify the data loading and processing scheme to support learning part-level concepts from single-image examples. In addition, since we can only access 24G VRAM GPU (RTX3090), we add LoRA [8] following standard StableDiffusion v2.1 model finetuning scheme to train the LoRA weights instead of the entire diffusion U-Net. This approach has been validated by PartCraft [14] that it will not harm the concept learning quality when compared to finetuning the entire U-Net.

PartCraft [14] is originally designed to learn part-level concepts from a large dataset of images. Specifically, they

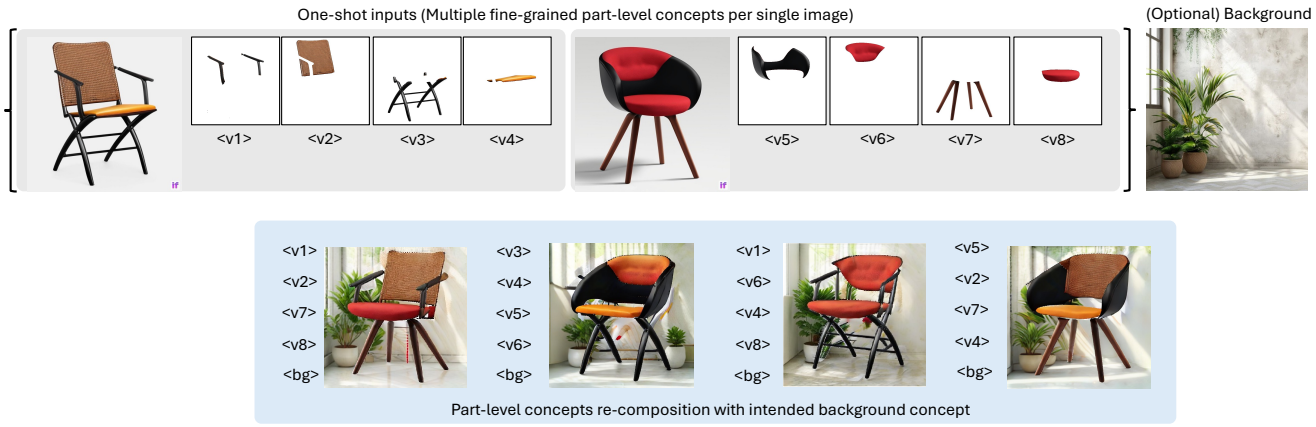

Figure 6. Concept mixing results for a pair of chairs with an intended background.

require to have 10-20 images for a subject (e.g., a specific bird species.). We modify the data loading and processing scheme to load single-image examples dataset into their pipeline. In addition, we directly provide the part-level masks instead of using their automatic concept discovering method to ensure fair comparison. We use StableDiffusion v1.5 as described in their paper to learn the concepts. Their approach show unstable performance. The results we show in Figure 5 are already the best results we observed. The image quality degrades quickly when the training step number increases (after the step we shown in Figure 5). We also tried to use smaller learning rate but their pipeline struggle to even learn any meaningful concepts. We hypothesis the reason for this is that their method is designed to use on a large dataset instead of the single-image examples in our case.

F. Extra Experimental Results

F.1. Intra-category Results

We conduct experiments on various object categories to demonstrate our model's capability in learning and recomposing part-level concepts. In each case, objects are decomposed into over four semantic parts, and our method learns to mix these parts across instances from the same category.

Chair. Parts are: `armrest`, `backrest`, `legs`, and `seat`. Figure 7 and Figure 8 show concept mixing results across 8 different chair pairs.

Bed. Parts are: `headboard`, `base`, `mattress`, and `pillow`. Figure 9 shows concept mixing for a pair of beds.

Gym Equipment. Parts are: `base`, `stand`, `seat`, and `weight`. Figure 10 shows concept mixing for a pair of gym equipment examples.

Vehicle. Parts are: `front`, `cockpit`, `tail`, and `wheels`. Figure 11 shows concept mixing for three pairs of vehicles.

Bike. Parts are: `handle`, `frame`, `seat`, and `wheels`. Figure 12 shows concept mixing results for a pair of bikes.

Plane. Parts are: `body`, `wing`, `tail`, and `engine`. Figure 13 shows concept mixing results for a pair of planes.

Bird and Virtual Creature. Parts are: `head`, `body`, `wings`, `tail`, and `legs`. Figure 14 shows concept mixing for a bird and a creature pair.

Virtual Characters. Figure 15 shows concept mixing for a mushroom-like character and a santa-like character. Parts are: `head`, `eyes`, `face`, `body`, and `legs`. Figure 16 shows concept mixing for a hermit-like character and a reptile-like character. This example aim to demonstrate that our method can learn a large number (i.e., in this case, 7 parts per image) of fine-grained concepts for parts and recompose them. Parts are: `head`, `eyes`, `face`, `nose`, `arms`, `body`, and `legs`.

F.2. Cross-category Results

To demonstrate generalization beyond intra-category recomposition, we evaluate our method on cross-category part mixing. Figure 17 shows 2 hybrid compositions, one for a chair and a gym equipment and another for a chair and a bed.

Figure 7. Concept mixing results for various compositions of chair parts.

G. Limitations

Although our method can learn well entangled part-level concepts can re-compose them, the image generation quality for some challenging inputs might still need improvement. For example, when composing parts from a truck and s sports car in Figure 11, there are some noticable artifacts in the generated images even though all the intended part concepts are preserved. Our pipeline also sometimes do not preserve the exact details for very thing or high-frequency structures like the horizontal bars in chair legs (in Figure 5) and the number of bird legs (in Figure 14). This problem might partially due to the inherent limitation of text-to-image diffusion models since these models can produce images containing unrealistic details like a bird with 3 or more legs. Additionally, for cross-category part re-composition, not all compositions of parts result in meaningful virtual objects. For example, when mixing a chair and a gym equipment in Figure 17, the composition in the third column results in non-meaningful objects. A composition prediction scheme may be developed in this case to predict the possible meaningful part compositions instead of naively trying all possible compositions.

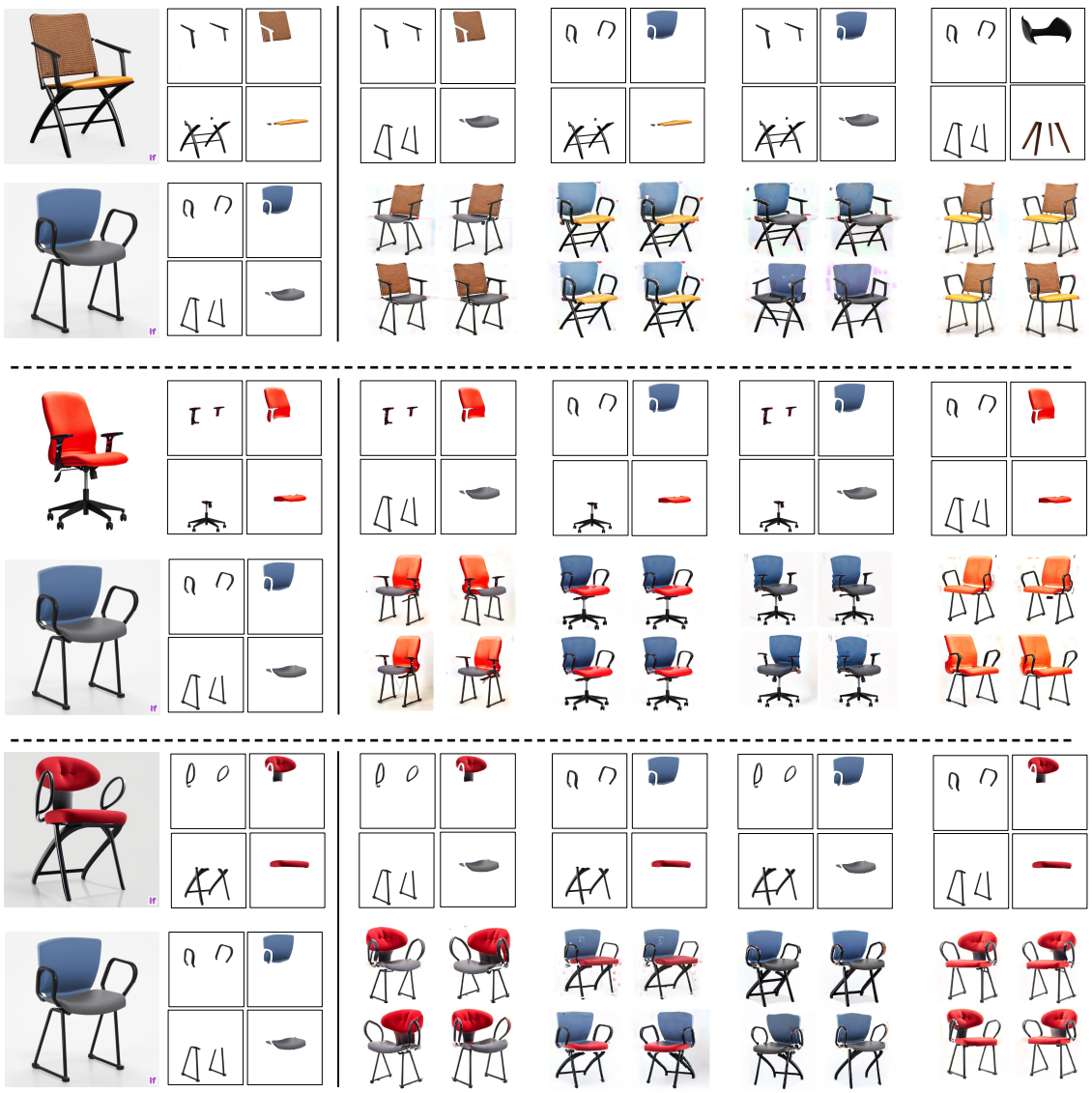

Figure 8. Concept mixing results for various compositions of chair parts.

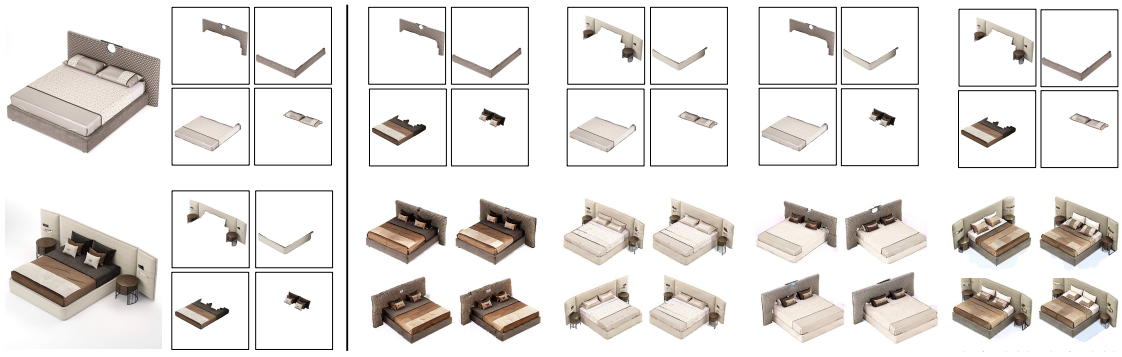

Figure 9. Concept mixing results for bed.

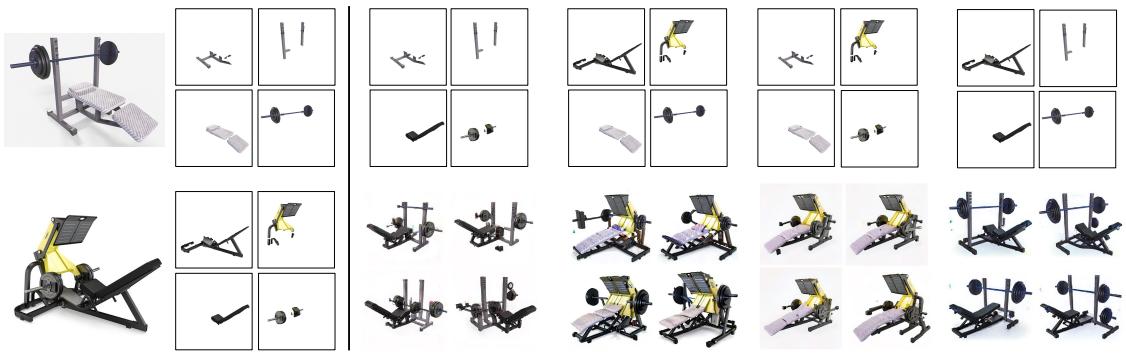

Figure 10. Concept mixing results for gym.

Figure 11. Concept mixing results for vehicles. The first pair contains a sports car and an old formula one car. The second pair contains a sports car and a modern formula one car. The last pair contains a sports car and a truck.

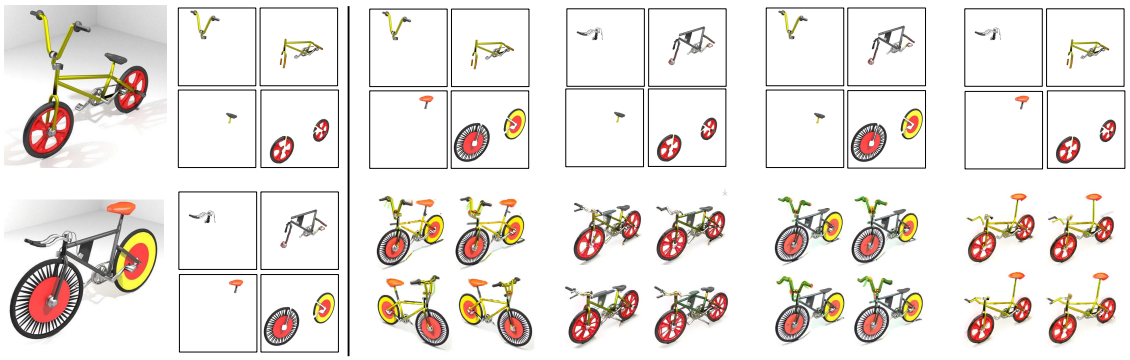

Figure 12. Concept mixing results for bike.

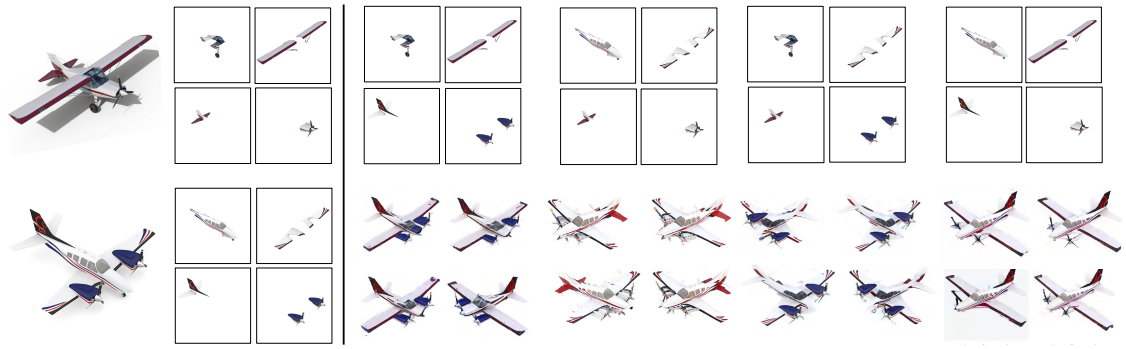

Figure 13. Concept mixing results for plane.

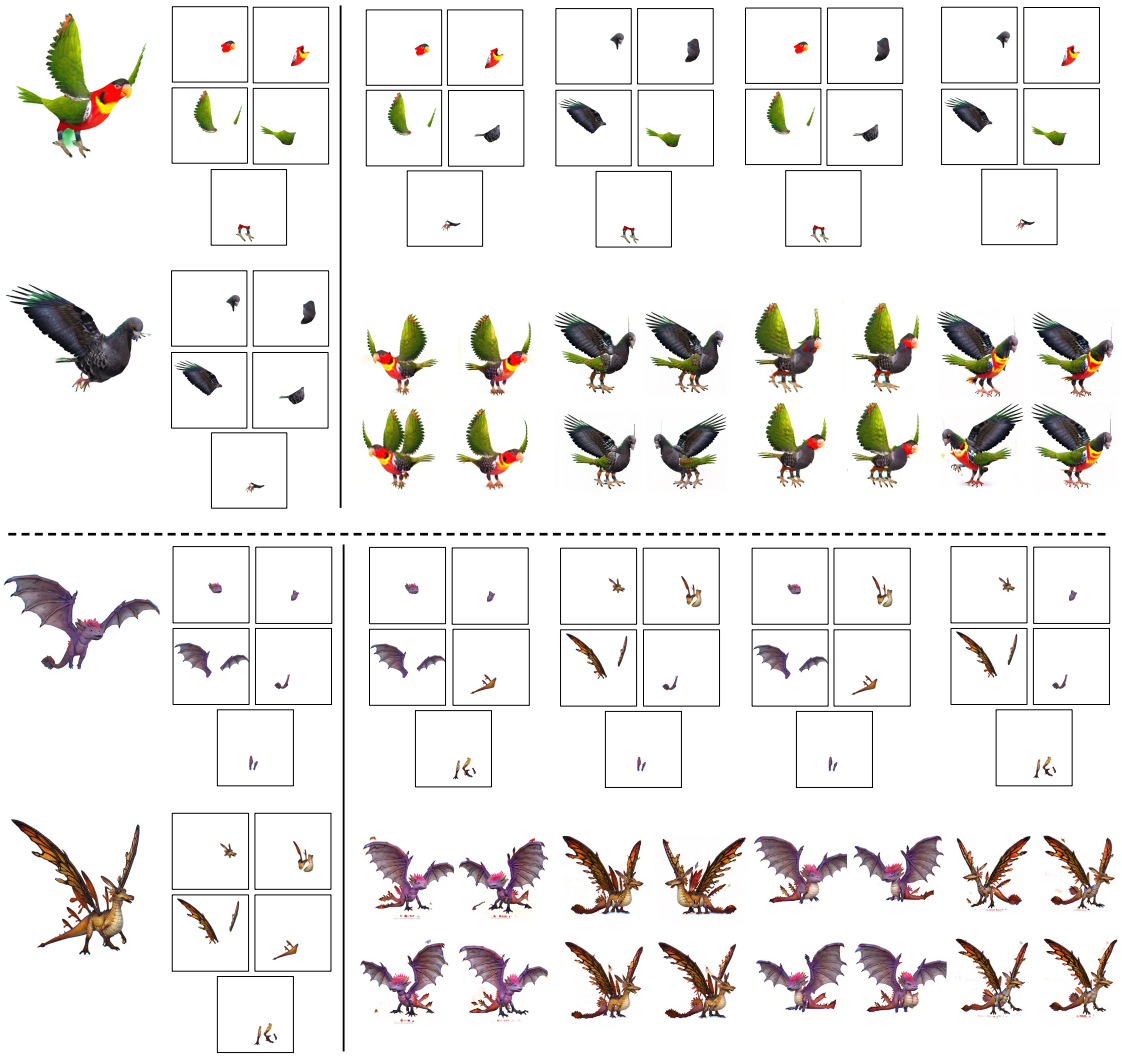

Figure 14. Concept mixing results for bird and virtual creatures.

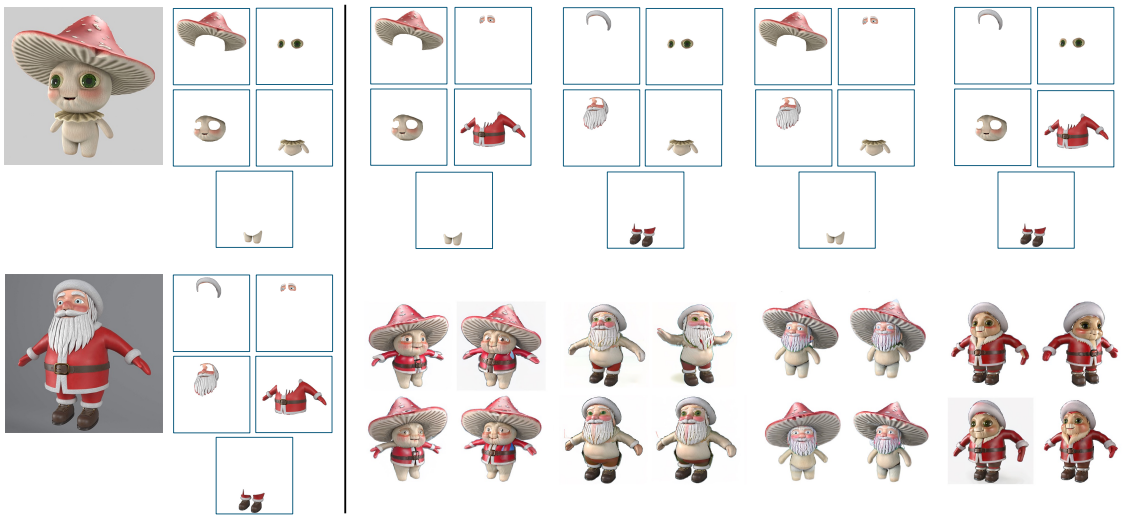

Figure 15. Concept mixing results for a mushroom-like character and a santa-like character .

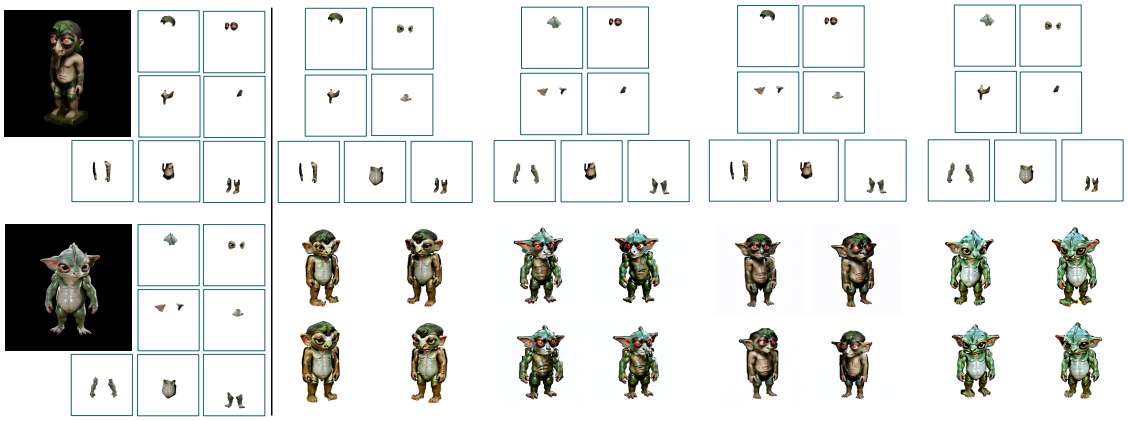

Figure 16. Concept mixing results for a hermit-like character and a reptile-like character.

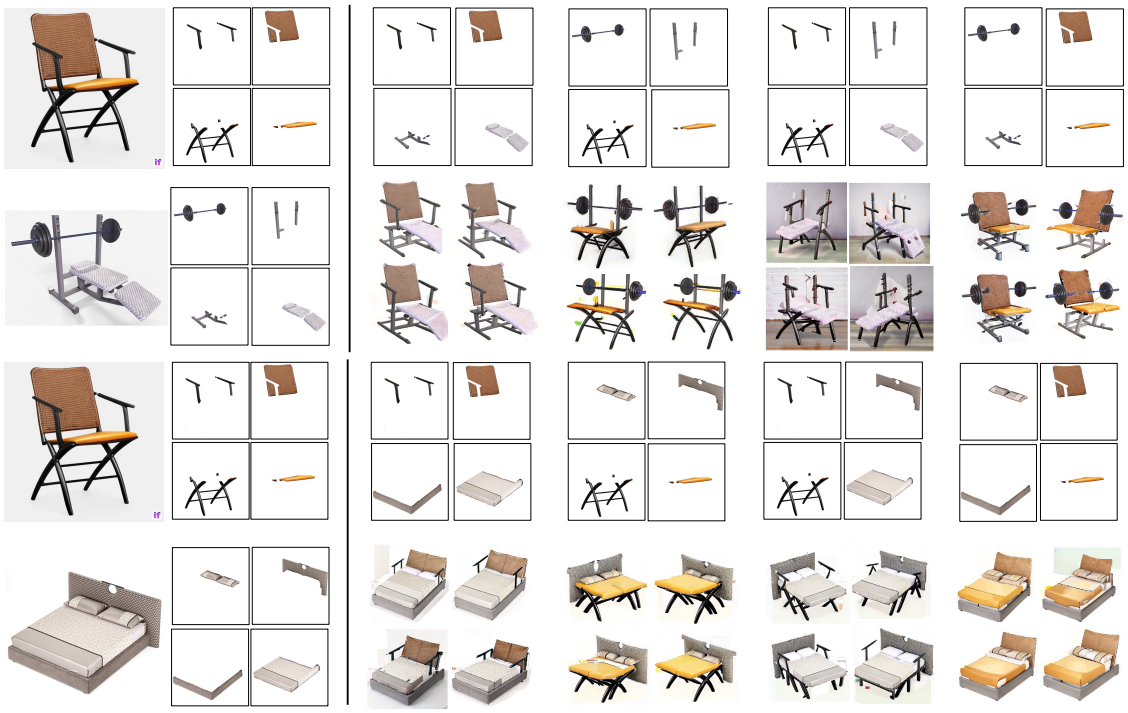

Figure 17. Concept mixing results cross-category objects. The first row shows the re-composing results for a chair and a gym equipment. The second row shows results for re-composing for a chair and a bed.