PartCraft: Crafting Creative Objects by Parts

Kam Woh Ng[1], Xiatian Zhu[1,2], Yi-Zhe Song[1], and Tao Xiang[1]

[1] CVSSP, University of Surrey, United Kingdom
[2] Surrey Institute for People-Centred AI
{kamwoh.ng,xiatian.zhu,y.song,t.xiang}@surrey.ac.uk

Abstract. This paper propels creative control in generative visual AI by allowing users to "select". Departing from traditional text or sketch-based methods, we for the first time allow users to choose visual concepts by parts for their creative endeavors. The outcome is fine-grained generation that precisely captures selected visual concepts, ensuring a holistically faithful and plausible result. To achieve this, we first parse objects into parts through unsupervised feature clustering. Then, we encode parts into text tokens and introduce an entropy-based normalized attention loss that operates on them. This loss design enables our model to learn generic prior topology knowledge about object's part composition, and further generalize to novel part compositions to ensure the generation looks holistically faithful. Lastly, we employ a bottleneck encoder to project the part tokens. This not only enhances fidelity but also accelerates learning, by leveraging shared knowledge and facilitating information exchange among instances. Visual results in the paper and supplementary material showcase the compelling power of PartCraft in crafting highly customized, innovative creations, exemplified by the "charming" and creative birds in Fig. 1. Code is released at `https://github.com/kamwoh/partcraft`.

Keywords: Part Composition · Controllable Text-to-image Generation

1 Introduction

Humans are creators; AI, on the other hand, hallucinates. Creativity, arguably, is the very force driving humanity forward. Recently, generative AI has garnered considerable attention for its perceived "creativity" [6, 19, 20, 43, 47, 49, 52, 55]. Despite its promise, the challenge of control has swiftly surfaced – how can humans infuse their creativity into the generation process and regulate the extent to which AI hallucinates?

Creativity starts with an idea. The immediate challenge is how to articulate that idea and integrate it with generative AI. Text is the most commonly employed medium. For instance, imagine being a bird enthusiast wanting to craft the most unique bird akin to Fig. 1. The go-to approach would be furnishing Stable Diffusion [52] (or an equivalent model) with the following textual prompt (the idea): *"generate a bird with the head of X, wings of Y, and body of Z".* While you might be presented with remarkably looking birds like Fig. 1(a), they

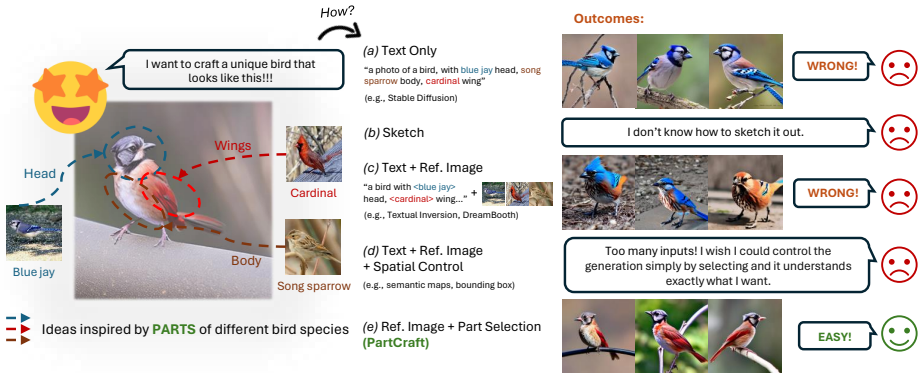


Fig. 1: People often form creative concepts based on existing ones [9, 42, 54, 65]. For instance, a bird enthusiast may want to craft a unique bird with different parts (*e.g.*, heads, bodies, and wings) from common bird types (*e.g.*, blue jay, cardinal and song sparrow). **(a)** Using text prompts in T2I models often results in a lack of control and deviation from the intended details, especially those visual details that are difficult to describe. **(b)** While sketching is a direct way, not everyone possesses the ability to sketch, particularly in intricate detail. **(c)** Even with reference images, existing methods (*e.g.*, DreamBooth [53]) did not consider learning object parts, thus cannot generate with desired parts. **(d)** Using additional control is even cumbersome, requiring too many inputs! **(e)** We aim to create an object by simply selecting desired parts. PartCraft learns from visual examples to generate the object with a faithful holistic structure, seamlessly integrating the chosen parts into a natural and coherent entity.

may bear little resemblance to the envisioned concept. Recent literature suggests that a swift sketch could serve as a viable alternative [40, 63, 71], providing fine-grained shape control. However, the caveat is that not everyone possesses the ability to sketch, particularly in intricate detail.

In this paper, our primary focus revolves around addressing the issue of "control" in generative AI. We advance by introducing fine-grained control into the generative process, inviting you to "select". While this selection mechanism might seem modest at first, it closely mirrors the human creative process, where new concepts often emerge from existing ones [9,42,54,65]. Recall those moments when you desired an "ideal" pair of shoes with selected features from different pairs, or when you aimed to get creative with a cat (for that matter!)?

It follows that rather than relying on writing (text) or drawing (sketch), all that is required is to choose the distinct visual concepts you specifically desire in your creative endeavor. Our model then ensures that all selected concepts are seamlessly and precisely composed into a faithful novel object in the final generation, without resorting to additional control such as bounding boxes [7, 38, 40, 66, 71]. To illustrate with the "unique bird" example once more, our approach literally involves selecting the head of X, wings of Y, and body of Z! (see Fig. 1(e) and 2).

Our solution is intuitive and centers around the well-studied computer vision concept of objects and their parts [8,14,22,27,35]. The challenge then boils down

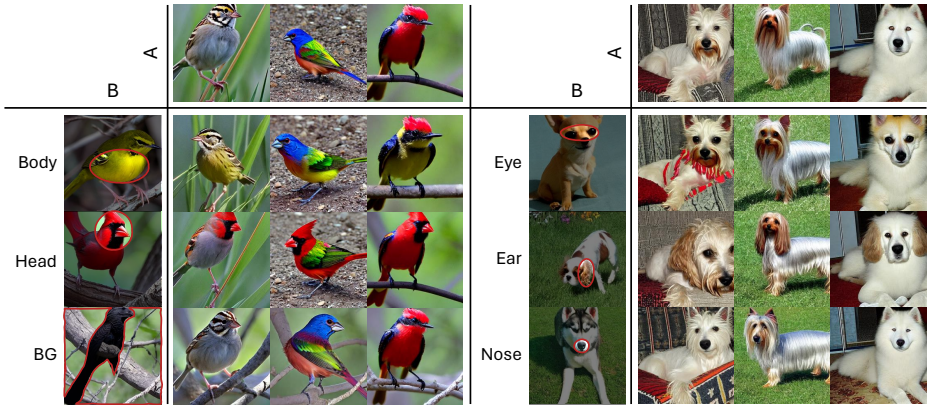


Fig. 2: Two sets of images were generated from their original parts (sources A and B). We can integrate a specific part (*e.g.*, body, head, or even background (BG) of source B to target A seamlessly without effort.

to two aspects: (i) how to parse known objects into their parts (*i.e.*, recognizing that birds have eyes and tails), and (ii) how to assemble parts from different visual concepts to form a faithful creative concept (*i.e.*, ensuring the output is recognizably a bird).

The former is easier. We make clever use of DINOv2 [45] feature maps and perform unsupervised clustering to identify common parts. The idea is that each cluster will then correspond to a semantic part of a common object (*e.g.*, head and wings of birds). We specifically chose DINOv2 for its superior fine-grained perception [2] compared to others [50, 57, 58]. Further, this way has a higher flexibility to enhance fine-grained parsing by using a higher cluster count.

Our major contribution lies in addressing the latter challenge. The solution is intuitive – it essentially revolves around the fine-grained selection and placement of the chosen parts. Inspired by recent efforts in personalization [3, 25, 36, 53, 64], primarily designed for learning entire objects, we introduce a tailored attention loss that specifically operates on parts. With this loss, our model learns the object part composition, ensuring the final generation appears holistically faithful with mixed parts (*i.e.*, head and wings of a bird actually appear at the right places).

More specifically, we introduce an entropy-based attention loss that maximizes the attention of a specific part at a particular location while minimizing the attention where no parts appear. This is achieved by first selecting the attention maps that correspond to the parts. A normalization over all parts is then performed to ensure that each image region is occupied by no more than one part. Finally, we minimize the entropy loss between each normalized individual part and the semantic maps obtained during object parsing, containing the correct part location. This not only facilitates stronger part disentanglement but is also the key to generating a faithful holistic structure of an object as it learns a generic prior topology knowledge about object parts.

To enhance generation fidelity, we further employ a bottleneck encoder to project the text tokens. This approach accelerates learning by leveraging shared knowledge (common parts) and facilitating information exchange among instances in each part. Each instance adjusts slightly to adapt to the fine-grained part details during optimization.

Our contributions are as follows: **(i)** We pioneer a unique approach for fine-grained part-level control in Text-to-Image (T2I) models, empowering users to craft objects by selecting desired parts. This method streamlines the creative process, marking a significant advancement in our capacity to manipulate and reimagine visual content. **(ii)** We introduce **PartCraft**, a technique that autonomously parses object parts and orchestrates them from different visual concepts, resulting in the faithful creation of a novel object. **(iii)** For enhanced part disentanglement and generation fidelity, we propose an entropy-based normalized attention loss and leverage a bottleneck encoder. **(iv)** We present two problem-specific quantitative metrics. Comprehensive experiments on CUB-200-2011 (birds) and the Stanford Dogs dataset demonstrate the superior performance of our method in generating novel objects, surpassing alternative approaches in both qualitative and quantitative evaluations.

2 Related Work

Creative editing and generation. Creativity involves generating innovative ideas or artifacts across various domains [12]. Extensive research has explored the integration of creativity into Generative Adversarial Networks (GANs) [21,44,56] and Variational Autoencoders (VAEs) [16, 18]. For example, DoodlerGAN [26] learns and combines fine-level part components to create sketches of new species. A recent study by [60] demonstrated decomposing personalized concepts into distinct visual aspects, creatively recombined through diffusion models. InstructPix2Pix [10] allows creative image editing through instructions, while ConceptLab [51] aims to identify novel concepts within a specified category, deviating from existing concepts. Different from these works where editing/generation usually focuses on the whole object, we instead focus on training a text-to-image generative model that can understand parts, thus able to creatively generate new objects by seamlessly composing different parts simply through selection.

Text-to-image generation. Recent advancements in large text-to-image (T2I) diffusion models [6, 19, 20, 43, 49, 52, 55] have made significant improvements over conventional methods [37, 46, 59, 67, 70, 72] in producing high-fidelity images from text prompts. Their application scope has expanded to include both global [10, 32, 41] and localized [4, 17, 28, 68] image editing tasks, demonstrating versatility. Methods such as [5, 7, 13, 15, 23, 24, 34, 38, 66, 69, 71] have introduced more granular spatial control, such as incorporating semantic segmentation masks or bounding boxes, into large pretrained diffusion models to guide image generation. Contrary to these approaches, we focus on enhancing the control by enabling a straightforward discrete selection of desired parts. We minimize the complexity and manual intervention required by spatial controls, yet the model can compose selected parts as a coherent object seamlessly. This not

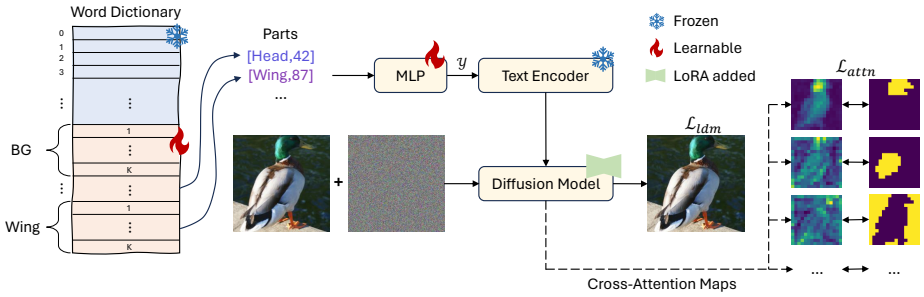


Fig. 3: Overview of our PartCraft. All parts are organized into a dictionary, and their semantic embeddings are learned through a textual inversion approach. For instance, a text description like "a photo of a [Head,42] [Wing,87]..." guides the optimization of the corresponding textual embedding by reconstructing the associated image. To improve generation fidelity, we incorporate a bottleneck encoder f (MLP) to compute the embedding y (Eq. (4)) as input to the text encoder. To promote disentanglement among learned parts, we minimize a specially designed attention loss, denoted as $\mathcal{L}_{attn}$.

only simplifies the user's role in the generative process but also ensures that the compositional logic and coherence are inherently managed by the model's ability. As such, our research is distinguished by focusing on the model's inherent ability to understand and apply part relationships instead of providing additional controls.

Abstracting visual knowledge as a text token. The effectiveness of T2I models is constrained by the user's ability to articulate their desired image through text. These models face challenges in faithfully replicating visual characteristics from a reference set and generating innovative interpretations in diverse contexts, even with detailed textual descriptions. To address this challenge, various personalization techniques have been developed. These techniques obtain a new word embedding from multiple images depicting the same concept [1, 25, 36, 53, 61] or multiple new word embeddings for various concepts within a single image [3] through inversion. The learned visual concepts can be creatively reused in many image editing tasks. Nonetheless, most of these approaches struggle to learn object parts, often not able to follow the part selections due to part entanglement as they were designed to learn object as a whole. In this work, we introduce a customized attention loss that serves a dual purpose: ensure accurate positioning of each part and enforce each image region occupied by no more than one part. This greatly improves the part disentanglement, creating novel concepts with correct appearances (see Fig. 2).

3 Methodology

Given a set of unlabeled images depicting the same object (*e.g.*, bird) with different part details, we aim to train a T2I generative model that decomposes

parts of objects into text tokens and can recompose them in a novel way. To that end, we propose PartCraft, as depicted in Fig. 3.

We start by discovering the parts in a three-tier hierarchy, as detailed in Sec. 3.1. Paired with the training images $\{x_i\}_{i=1}^N$, this semantic hierarchy subsequently serves as the supervision to fine-tune a pre-trained text-to-image model, say a latent diffusion model [52], denoted as $\{\epsilon_\theta, \tau_\theta, \mathcal{E}, \mathcal{D}\}$, where ϵ_θ represents the diffusion denoiser, τ_θ the text encoder, and $\mathcal{E}/\mathcal{D}$ the autoencoder respectively. We adopt the textual inversion technique [25]. Concretely, we learn a set of pseudo-words p^* for each part in the word embedding space with:

$$\mathcal{L}_{ldm} = \mathbb{E}_{z,t,p,\epsilon}\big[||\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y_p))||_2^2\big], \tag{1}$$

$$p^* = \underset{p}{\text{argmin}}\ \mathcal{L}_{ldm}, \tag{2}$$

where $\epsilon \sim \mathcal{N}(0,1)$ denotes the unscaled noise, t is the time step, $z = \mathcal{E}(x)$ is the latent representation of the image, z_t is the latent noise at time t, and y_p is the text condition that includes p as part of the text tokens. $\mathcal{L}_{ldm}$ is a standard diffusion loss [30] to reconstruct the parts. As each object is composed of a set of parts, its reconstruction is achieved by the reconstruction of the associated set of parts. In other words, when all parts are reconstructed properly, it will become a valid object.

3.1 Unsupervised Part Discovery

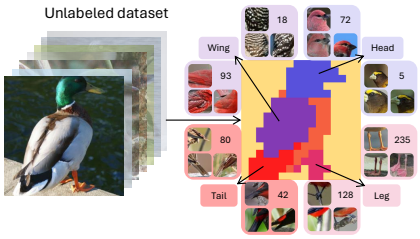


Fig. 4: Part discovery within a semantic hierarchy involves partitioning each image into distinct parts and forming semantic clusters across unlabeled training data.

To minimize the labeling cost, we develop a scalable process to reveal the underlying semantic hierarchy with parts in an unsupervised fashion. We leverage the off-the-shelf vision model for image decomposition and clustering. Specifically, given an image x_i, we employ DINOv2 [45] to extract the feature map $F = \{F_i = \text{DINOv2}(x_i)\}_i^N$. We then conduct three-level hierarchical clustering (see Fig. 4): **(i)** At the top level, k-means is applied on all patches in F with $k = 2$ to separate foregrounds and backgrounds B. **(ii)** At the middle level, k-means is further applied on all foreground patches to acquire M clusters representing common parts, such as the heads of birds. **(iii)** At the bottom level, we further group each of the M clusters as well as the background cluster B into K splits. Each split refers to finer meanings, such as the head of a specific bird species, or a specific background style. Lastly, each region of an image will be tagged with the corresponding cluster index. We represent these cluster tags as follows:

$$p = (0, k_0), (1, k_1), ..., (M, k_M), \tag{3}$$

where the first pair refers to the background style, and the following M pairs denote the combinations of M parts (*e.g.*, head, body, wings) each associated with a specific object (*e.g.*, sparrow), and $k \in \{1, \ldots, K\}$. This description will be used as the textual prompt in model training, such as "a photo of a [p]". Please refer to the supplementary material for more examples of the discovered semantic hierarchy. This process also yields the segmentation mask of each m-th part, which we define as S_m.

Motivation. While we can leverage off-the-shelf segmentation models such as VLPart [58], the robustness relies on the generalizability of the model and the part segmentation result is usually pre-defined and may be unstable for unseen domains. As a result, we rely on our feature clustering method to obtain the segmentation map, which also has a higher flexibility in choosing the number of clusters (parts).

3.2 Part Token Bottleneck

In contrast to prior text inversion studies [25], our task requires learning a greater quantity at the same time – specifically, $(M+1)K$-of word tokens derived from a collection of discovered parts marked by inherent imperfections (such as partial overlap and over splitting). This makes the learning task more demanding. To enhance the learning process, we propose a neural network $f(\cdot)$ comprising a two-layer MLP with ReLU activation:

$$y_p = f(e(p)), \tag{4}$$

where y_p will be subsequently used as the input[3] to the text encoder τ_θ and $e \in \mathbb{R}^{MK \times D}$ is a learnable word embedding dictionary that maps p to their respective embeddings.

Our design demonstrates quicker convergence than directly learning the final word embeddings $e(\cdot)$ [25] (see supplementary material). This could be attributed to the entanglement of word embeddings in the conventional design, where there is no information exchange among them during optimization. For instance, each token doesn't know they are learning for a specific part of a specific species. This lack of communication leads to lower data efficiency and slower learning. With the bottleneck f, it will first project the token into a common part embedding space (*e.g.*, head), then slightly adjust itself to adapt the fine-grained part details. It's worth noting that the conventional design is a specific instance of our approach when f is an identity function.

3.3 Learning to Craft by Parts

Fine-tuning the T2I model, rather than sorely learning pseudo-words, has been shown to achieve better reconstruction of target concepts as demonstrated in [36, 53]. However, this comes with a significant training cost. Thus, we apply LoRA

[3] Word templates such as "a photo of a [*]" will be used.

(low-rank adaptation) [31] to the cross-attention block for efficient training. We then minimize the diffusion loss $\mathcal{L}_{ldm}$ (Eq. (1)) to learn both pseudo-words and LoRA adapters.

While training with only $\mathcal{L}_{ldm}$, entanglement happens between parts, as evident from the attention maps in the cross-attention block of the denoiser ϵ_θ (see Fig. 9). This entanglement arises due to the correlation between parts (*e.g.*, a bird head code is consistently paired with a bird body code to represent the same species). To address this issue, we introduce an entropy-based attention loss as regularization:

$$\mathcal{L}_{attn} = \mathbb{E}_{z,t,m}\big[-\big(S_m \log \hat{A}_m + (1 - S_m)\log(1 - \hat{A}_m)\big)\big], \tag{5}$$

$$\hat{A}_{m,i,j} = \frac{\bar{A}_{m,i,j}}{\sum_k \bar{A}_{k,i,j}}, \quad \bar{A}_m = \frac{1}{L}\sum_l^L A_{l,m}, \tag{6}$$

where $A \in [0,1]^{M \times HW}$ represents the cross-attention map between the m-th part and the noisy latent z_t, L represents the number of specific layers to select attention maps, $\hat{A} \in [0,1]^{M \times HW}$ represents the averaged and normalized cross-attention map over all parts and $S_m \in \{0,1\}^{M \times HW}$ serves as the mask that indicates the location of m-th part. In cases where the part is not present in the image (*e.g.*, occluded), we set both S_m and $\hat{A}_m$ as 0 to exclude them. Thus, the overall learning objective is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{ldm} + \lambda_{attn}\mathcal{L}_{attn}, \tag{7}$$

where $\lambda_{attn} = 0.01$. We focus on attention maps at the resolution of 16×16 where rich semantic information is captured [28]. Normalization is performed at each location to ensure that the sum of a patch location equals 1. This aims to maximize the attention of a specific part at a particular location which implicitly minimizing the attention of other parts similar to a softmax classification task. Compared to the mean-square based attention loss [3], this intuitively ensures that a part only appears once at a particular location, facilitating stronger disentanglement from other parts during the denoising operation. When generating a part for a particular location, the diffusion model ϵ_θ should only attend to the part instead of other non-related parts.

4 Experiments

Datasets. We demonstrate our selection task on two fine-grained object datasets: CUB-200-2011 (birds) [62] which contains 5,994 training images, and Stanford Dogs [33] which contains 12,000 training images.

Implementation. For part composition, we assess the model's ability to combine up to 4 different parts from 4 distinct species/objects. We set $M = 5$ for bird generation (head, front body/breast area, wings, legs, tail) and $M = 7$ for dog generation (forehead, eyes, mouth/nose, ears, neck, body/tail, legs). For both datasets, K is set as 256, ensuring sufficient coverage of all fine-grained

Method	Learnable Token	Fine Tune	Disentanglement	Bottleneck
Textual Inversion [25]	✓	✗	✗	✗
DreamBooth [53]	✓	LoRA*	✗	✗
CustomDiffusion [36]	✓	K/V	✗	✗
Break-a-scene [3]	✓	LoRA*	MSE	✗
PartCraft (Ours)	✓	LoRA	Eq. (5)	✓

Table 1: Comparing our and alternative methods in design properties. *: We fine-tuned the added LoRA [31] adapter rather than the entire diffusion model ϵ_θ due to resource limit. MSE is a mean-square based attention loss used in [3].

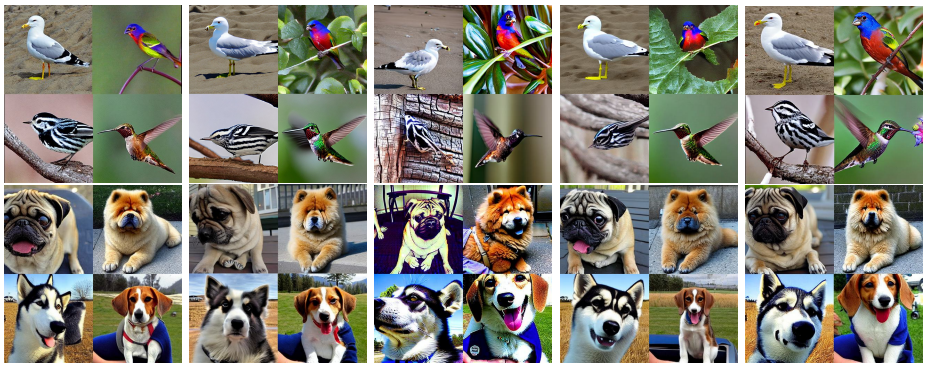
classes (*i.e.*, 200 for birds and 120 for dogs). We randomly generate 500 images by sampling 500 sets of parts. For each image, we randomly replace an original part with any part from another 500 non-overlapping sets of parts. The resulting set of parts may take the form of "$(0, k_A)$ $(1, k_B)$ $(2, k_C)$ ... (M, k_D)", representing a composition from species A, B, C, and D. Stable Diffusion v1.5 [52] is used. Please see the supplementary material for further training details.

We compare our method with the recent personalization methods: Textual Inversion (TI) [25], DreamBooth (DB) [53], Custom Diffusion (CD) [36], Break-a-scene (BaS) [3]. These personalization methods were designed to take single or multiple images with associated labeled objects as input. The text prompt for each image is as simple as "a photo of [p]" where p is expressed in Eq. (3), since we do not rely on complex prompts. We employ the official implementations released by the original authors for training. We summarize the main design properties of all compared methods in Tab. 1.

Evaluation metrics. To assess a model's ability to disentangle and composite parts, we introduce two metrics: (a) exact matching rate (*EMR*) and (b) cosine similarity (*CoSim*) between the k-means embeddings of the parts of real and generated images. Utilizing the pre-trained k-means from Sec. 3.1, we predict the parts of generated images. *EMR* quantifies how accurately the cluster index of parts of generated images matches the parts of the corresponding real images whereas *CoSim* measures the cosine similarity between the k-means centroid vector that the part belongs to between generated and real images. These metrics assess the model's ability to follow the input parts and accurately reconstruct them, with perfect disentanglement indicated by *EMR* of 1 and *CoSim* of 1. A detailed algorithm is provided in the supplementary material. We also measure image generation quality using FID [29] to assess model performance in terms of image distribution. Additionally, we compute the average pairwise cosine similarity between CLIP [48]/DINO [11] embeddings of generated and real class-specific images following [53]. Each generated image is conditioned on the parts of the corresponding real image. This results in 5,994 generated images for birds and 12,000 generated images for dogs.

Method	Birds: CUB-200-2011					Dogs: Stanford Dogs				
	FID	CLIP	DINO	EMR	CoSim	FID	CLIP	DINO	EMR	CoSim
Textual Inversion [25]	**10.10**	**0.784**	0.607	0.305	0.842	23.36	0.652	0.532	0.218	0.754
DreamBooth [53]	12.94	0.775	0.594	0.355	0.856	22.65	0.660	0.563	0.275	0.777
Custom Diffusion [36]	37.61	0.694	0.504	0.338	0.833	42.41	0.593	0.491	0.253	0.755
Break-a-Scene [3]	20.05	0.742	0.549	0.390	0.854	24.20	0.633	0.532	0.300	0.775
PartCraft (Ours)	12.86	0.783	**0.618**	**0.460**	**0.882**	**16.92**	**0.669**	**0.573**	**0.358**	**0.796**

Table 2: Quantitative comparison for part reconstruction.



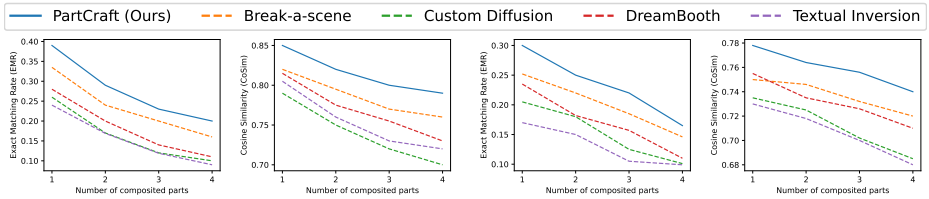
(a) TI **(b)** DB **(c)** CD **(d)** BaS **(e)** PartCraft

Fig. 5: Visual comparison under the part reconstruction setting. All images are generated by using the original parts of respective objects.

4.1 Part Reconstruction and Image Quality Evaluation

We first assess the ability of different methods to learn parts as text tokens by evaluating how well they can accurately reconstruct the parts (this also means image generation with original parts).

In Tab. 2, we summarize the performance of respective methods on the bird and dog generation, respectively. We highlight four observations: **(i)** Textual Inversion performs quite well compared to DreamBooth, CustomDiffusion, and Break-a-scene in terms of FID, CLIP, and DINO scores although did not fine-tune the diffusion model ϵ_θ. This may be due to the potential risk of overfitting when fine-tuning ϵ_θ especially when learning a vast array of new concepts with many update iterations. It is also not uncommon to carefully tune the learning rate and the training iterations in these models when fine-tuning new concepts (*e.g.*, only 800-1000 steps of updates to learn a new concept in [3]). **(ii)** Nonetheless, fine-tuning the diffusion model ϵ_θ can help improve the ability to follow prompts as shown by increased EMR and CoSim scores (*e.g.*, EMR of at least 5% in DreamBooth). **(iii)** Break-a-scene has a better ability to reconstruct the parts as shown by EMR and CoSim, this is due to the attention loss explicitly forcing the parts to focus on the respective semantic region. **(iv)** PartCraft achieves the best performance in DINO, EMR, and CoSim scores (*e.g.*, 7% better in EMR compared to Break-a-scene). This indicates that not only does our



(a) CUB-200-2011 **(b)** Stanford Dogs

Fig. 6: Quantitative comparisons of part composition in terms of EMR and CoSim.

image-generation ability perform comparably well with textual inversion, but PartCraft is also able to disentangle the parts learning so that it can follow the prompt instructions more accurately to generate the parts in a cohort. In Fig. 5, we present generated images from different methods, with CustomDiffusion exhibiting high-contrast images, possibly due to unconstrained fine-tuning on the cross-attention components K/V and resulting in worse FID scores.

4.2 Part Composition Evaluation

In this section, we assess the part composition ability of different methods. In this experiment, we generate the image by mixing different parts from different species. Our findings, as shown in Fig. 6, can be summarized as follows: **(i)** As the number of composited parts increases, EMR and CoSim decrease, reflecting the challenge of composing multiple diverse parts. **(ii)** Break-a-scene and PartCraft achieve notably higher EMR and CoSim scores, thanks to disentanglement through attention loss minimization. **(iii)** PartCraft outperforms Break-a-scene significantly by token bottleneck and tailored attention loss.

In Fig. 7, we visualize the results of composing 4 different parts. While all images appear realistic, most methods struggle to assemble all 4 parts. For instance, Break-a-scene missed out on the flurry body of *kerry blue terrier* (rightmost column). In contrast, our methods successfully combine 4 different parts from 4 different species, demonstrating the superior ability of our approach to part composition. We also visualize additional examples of our method in the supplementary material.

Furthermore, we explore the versatility of the adapted model by generating images with simple styles such as *pencil drawing*. While most methods successfully incorporate specific styles into the generated image, Custom Diffusion often fails to do so, possibly due to the unconstrained fine-tuning of the cross-attention components K/V.

4.3 Ablation Studies

Component analysis. In Fig. 8, we evaluate the effect of our proposed components (token bottleneck and attention loss) on creating novel bird species. **(i)** Removing the bottleneck outlined in Eq. (4) degrades the generation quality as evidenced by a higher FID score (12.86 $\rightarrow$ 16.36) even though both EMR

Fig. 7: Visual comparison on 4-species (specified on the top row) mixed generation. The last column indicates generated images with different styles (*i.e.*, *DSLR*, *Van Gogh*, *Oil Painting*, *Pencil Drawing*).

and CoSim remain. **(ii)** By replacing our $\mathcal{L}_{attn}$ with the MSE loss as proposed in [3], we observe significant deterioration in both EMR and CoSim. **(iii)** Finally, incorporating both the projector and our attention loss performs the best. This improvement highlights the necessity of incorporating interactions between multiple parts to achieve more effective part disentanglement and optimization.

Cross-attention visualization. Our attention loss plays a crucial role in token disentanglement. We demonstrate the impact of this loss in Fig. 9, where we

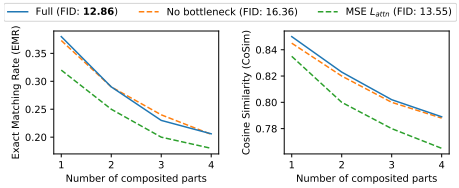


Fig. 8: Ablation on our part token bottleneck and attention loss under the part composition on CUB-200-2011 birds.

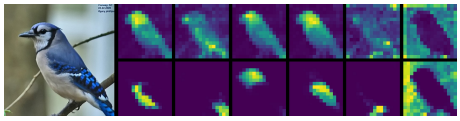
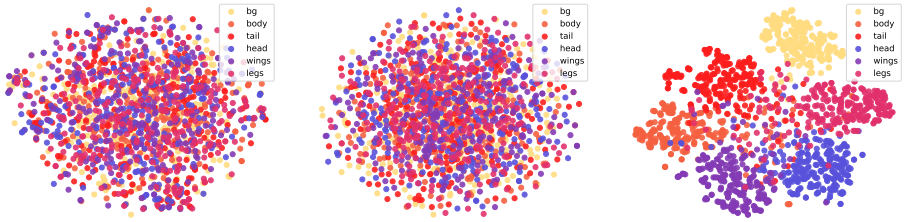


Fig. 9: Cross-attention map of each part (top) without and (bottom) with our attention loss.



(a) Textual Inversion **(b)** Break-a-scene **(c)** PartCraft (Ours)

Fig. 10: 2D tSNE [39] projection of word embeddings. Different colors represent different common parts. (correspond to the segmentation mask in Fig. 4).

observe significantly enhanced disentanglement after explicitly guiding attention to focus on distinct semantic regions.

Part word embedding space. We visualize the word embeddings of learned tokens of Textual Inversion [25], Break-a-scene [3] and our PartCraft for birds generation (CUB-200-2011 [62]) through tSNE [39] in Fig. 10. In our PartCraft, the word embeddings are the projected embeddings through Eq. (4). We can see that our projected version has a better semantic meaning such that the part embeddings are clustered together by their semantic meaning (*e.g.*, head). We believe this is one of the reasons that our PartCraft outperforms previous methods in which we can compose all parts seamlessly yet with higher quality.

Transferability for Creativity. (i) In Fig. 11, we demonstrate that not only it can compose parts within the training domain (*e.g.*, birds), but it can also transfer the learned parts to and combine with other domains (*e.g.*, cat). This enables the creation of unique combinations, such as a cat with a dog's ear. **(ii)** Leveraging the prior knowledge embedded in Stable Diffusion, PartCraft can also repurpose learned parts for creative image generation. An example of this is the generation of a bird-shaped robot adorned with various parts. These examples showcase PartCraft's immense potential for diverse and limitless creative applications. Please see the supplementary material for more examples.

Fig. 11: (Top): Using learned parts to modify the property of other domains such as cat, and lion with prompt such as "A cat with [*beagle's ear*]". **(Bottom)** We can also repurpose learned parts for creative image generation using prior knowledge in Stable Diffusion with prompt such as "A robot designed inspired by [*red header woodpecker's head*] and [*blue jay's body*]".

5 Conclusion

We propose a new way of control in generative AI. Instead of text or sketch, we "select" desired parts to create an object. We addressed the challenge of learning parts in T2I models by introducing a customized attention loss. This loss serves a dual purpose: to ensure parts are at the right location and to ensure each location is occupied by not more than one part. This greatly improves the part disentanglement. We further employ a non-linear bottleneck encoder to improve generation fidelity. Our model, PartCraft, can seamlessly compose different parts from different objects, creating objects that do not exist yet holistically correct and plausible objects by mixing them. Extensive experiments demonstrated PartCraft's superior performance in both qualitative and quantitative evaluation. Moreover, the learned parts demonstrate strong transferability. We hope that our PartCraft will empower artists, designers, and enthusiasts to bring the creations of their dreams to reality.

6 Limitations and Future Works

It is worth noting that the accuracy of obtained parts may be affected by using a self-supervised pre-trained feature extractor, incorporating encoders [64] may help. We also observed challenges in composing relatively small parts, like tails and legs. Additionally, we are also exploring cross-domain generation, *i.e.*, combining learned parts from different datasets to create objects with even more diverse parts. For instance, we can merge non-rigid parts (*e.g.*, parts from quadrupled animals) with rigid parts (*e.g.*, parts from cars/airplanes) and form a creative structure (*e.g.*, a car that has horse legs instead of wheels). This not only further improves the applicability of PartCraft but also serves as a stepping stone to creative generation, as generative AI progresses, continually expanding the limits of achievable creativity and artistic expression.

Acknowledgements

We extend our special thanks to Jia Wei Sii for her help in creating figures and discussing the main concept. We are also grateful to Ruoyi Du, Zhiyu Qu, Chee Seng Chan, and the reviewers for their fruitful comments and corrections on our draft, methodology, and experiments.

References

1. Alaluf, Y., Richardson, E., Metzer, G., Cohen-Or, D.: A neural space-time representation for text-to-image personalization. arXiv preprint arXiv:2305.15391 (2023)
2. Amir, S., Gandelsman, Y., Bagon, S., Dekel, T.: On the effectiveness of vit features as local semantic descriptors. In: ECCV (2022)
3. Avrahami, O., Aberman, K., Fried, O., Cohen-Or, D., Lischinski, D.: Break-a-scene: Extracting multiple concepts from a single image. In: SIGGRAPH Asia (2023)
4. Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. ACM TOG (2023)
5. Avrahami, O., Hayes, T., Gafni, O., Gupta, S., Taigman, Y., Parikh, D., Lischinski, D., Fried, O., Yin, X.: Spatext: Spatio-textual representation for controllable image generation. In: CVPR (2023)
6. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
7. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation. In: ICML (2023)
8. Biederman, I.: Recognition-by-components: a theory of human image understanding. Psychological review (1987)
9. Bonnardel, N., Marmèche, E.: Towards supporting evocation processes in creative design: A cognitive approach. International Journal of Human-Computer Studies (2005)
10. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR (2023)
11. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
12. Cetinic, E., She, J.: Understanding and creating art with ai: Review and outlook. ACM Transactions on Multimedia Computing, Communications, and Applications (2022)
13. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM TOG (2023)
14. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: deep learning for interpretable image recognition. NeurIPS (2019)
15. Chen, M., Laina, I., Vedaldi, A.: Training-free layout control with cross-attention guidance (2024)
16. Cintas, C., Das, P., Quanz, B., Tadesse, G.A., Speakman, S., Chen, P.Y.: Towards creativity characterization of generative models via group-based subset scanning. In: IJCAI (2022)
17. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. In: ICLR (2023)

18. Das, P., Quanz, B., Chen, P.Y., Ahn, J.w., Shah, D.: Toward a neuro-inspired creative decoder. In: IJCAI (2020)
19. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al.: Cogview: Mastering text-to-image generation via transformers. In: NeurIPS (2021)
20. Ding, M., Zheng, W., Hong, W., Tang, J.: Cogview2: Faster and better text-to-image generation via hierarchical transformers. In: NeurIPS. vol. 35 (2022)
21. Elgammal, A., Liu, B., Elhoseiny, M., Mazzone, M.: Can: Creative adversarial networks generating "art" by learning about styles and deviating from style norms. In: ICCC (2017)
22. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. IJCV (2005)
23. Feng, W., He, X., Fu, T.J., Jampani, V., Akula, A.R., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y.: Training-free structured diffusion guidance for compositional text-to-image synthesis. In: ICLR (2022)
24. Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., Taigman, Y.: Make-a-scene: Scene-based text-to-image generation with human priors. In: ECCV (2022)
25. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: ICLR (2023)
26. Ge, S., Goswami, V., Zitnick, C.L., Parikh, D.: Creative sketch generation. In: ICLR (2021)
27. He, J., Chen, J., Lin, M.X., Yu, Q., Yuille, A.L.: Compositor: Bottom-up clustering and compositing for robust part and object segmentation. In: CVPR (2023)
28. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. In: ICLR (2023)
29. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. NeurIPS **30** (2017)
30. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
31. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022)
32. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: CVPR (2023)
33. Khosla, A., Jayadevaprakash, N., Yao, B., Li, F.F.: Novel dataset for fine-grained image categorization: Stanford dogs. In: CVPRW (2011)
34. Kim, Y., Lee, J., Kim, J.H., Ha, J.W., Zhu, J.Y.: Dense text-to-image generation with attention modulation. In: ICCV (2023)
35. Krause, J., Jin, H., Yang, J., Fei-Fei, L.: Fine-grained recognition without part annotations. In: CVPR (2015)
36. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: CVPR (2023)
37. Li, B., Qi, X., Lukasiewicz, T., Torr, P.: Controllable text-to-image generation. In: NeurIPS (2019)
38. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: CVPR (2023)
39. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research (2008)

40. Mo, S., Mu, F., Lin, K.H., Liu, Y., Guan, B., Li, Y., Zhou, B.: Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. arXiv preprint arXiv:2312.07536 (2023)
41. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: CVPR (2023)
42. Nagai, Y., Taura, T., Mukai, F.: Concept blending and dissimilarity: factors for creative concept generation process. Design studies (2009)
43. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
44. Nobari, A.H., Rashad, M.F., Ahmed, F.: Creativegan: Editing generative adversarial networks for creative design synthesis. arXiv preprint arXiv:2103.06242 (2021)
45. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)
46. Qiao, T., Zhang, J., Xu, D., Tao, D.: Mirrorgan: Learning text-to-image generation by redescription. In: CVPR (2019)
47. Qu, Z., Xiang, T., Song, Y.Z.: Sketchdreamer: Interactive text-augmented creative sketch ideation. In: BMVC (2023)
48. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
49. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
50. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., Zhang, L.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024)
51. Richardson, E., Goldberg, K., Alaluf, Y., Cohen-Or, D.: Conceptlab: Creative generation using diffusion prior constraints. arXiv preprint arXiv:2308.02669 (2023)
52. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
53. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR (2023)
54. Runco, M.A., Jaeger, G.J.: The standard definition of creativity. Creativity research journal (2012)
55. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS. vol. 35 (2022)
56. Sbai, O., Elhoseiny, M., Bordes, A., LeCun, Y., Couprie, C.: Design: Design inspiration from generative networks. In: ECCVW (2019)
57. Sun, P., Chen, S., Luo, P.: Grounded segment anything: From objects to parts. `https://github.com/Cheems-Seminar/grounded-segment-any-parts` (2023)
58. Sun, P., Chen, S., Zhu, C., Xiao, F., Luo, P., Xie, S., Yan, Z.: Going denser with open-vocabulary part segmentation. arXiv preprint arXiv:2305.11173 (2023)

59. Tao, M., Tang, H., Wu, F., Jing, X.Y., Bao, B.K., Xu, C.: Df-gan: A simple and effective baseline for text-to-image synthesis. In: CVPR (2022)
60. Vinker, Y., Voynov, A., Cohen-Or, D., Shamir, A.: Concept decomposition for visual exploration and inspiration. In: SIGGRAPH Asia (2023)
61. Voynov, A., Chu, Q., Cohen-Or, D., Aberman, K.: P+: Extended textual conditioning in text-to-image generation. arXiv preprint arXiv:2303.09522 (2023)
62. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset (2011)
63. Wang, X., Darrell, T., Rambhatla, S.S., Girdhar, R., Misra, I.: Instancediffusion: Instance-level control for image generation. arXiv preprint arXiv:2402.03290 (2024)
64. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In: ICCV (2023)
65. Wilkenfeld, M.J., Ward, T.B.: Similarity and emergence in conceptual combination. Journal of Memory and Language (2001)
66. Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., Shou, M.Z.: Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In: ICCV (2023)
67. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: CVPR (2018)
68. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: CVPR (2023)
69. Yang, Z., Wang, J., Gan, Z., Li, L., Lin, K., Wu, C., Duan, N., Liu, Z., Liu, C., Zeng, M., et al.: Reco: Region-controlled text-to-image generation. In: CVPR (2023)
70. Yin, G., Liu, B., Sheng, L., Yu, N., Wang, X., Shao, J.: Semantics disentangling for text-to-image generation. In: CVPR (2019)
71. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
72. Zhu, M., Pan, P., Chen, W., Yang, Y.: Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In: CVPR (2019)