

# pop: Photo-Inspired Diffusion Operators

ANONYMOUS AUTHORS

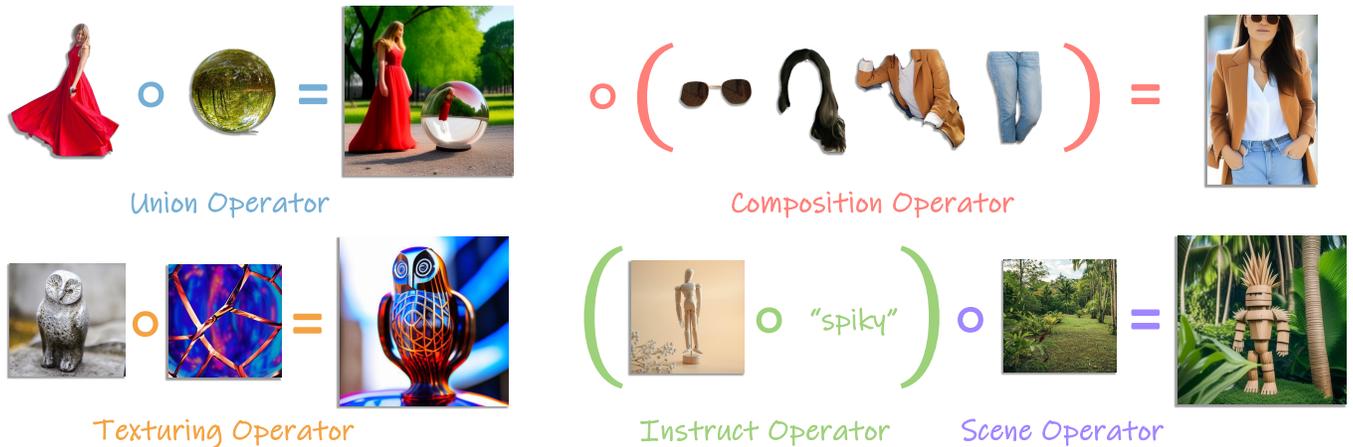


Fig. 1. Different operators trained using *pOps*. Our method learns operators that are applied directly in the image embedding space, resulting in a variety of semantic operations that can then be realized as images using an image diffusion model.

Text-guided image generation enables the creation of visual content from textual descriptions. However, certain visual concepts cannot be effectively conveyed through language alone. This has sparked a renewed interest in utilizing the CLIP image embedding space for more visually-oriented tasks through methods such as IP-Adapter. Interestingly, the CLIP image embedding space has been shown to be semantically meaningful, where linear operations within this space yield semantically meaningful results. Yet, the specific meaning of these operations can vary unpredictably across different images. To harness this potential, we introduce *pOps*, a framework that trains specific semantic operators directly on CLIP image embeddings. Each *pOps* operator is built upon a pretrained Diffusion Prior model. While the Diffusion Prior model was originally trained to map between text embeddings and image embeddings, we demonstrate that it can be tuned to accommodate new input conditions, resulting in a diffusion operator. Working directly over image embeddings not only improves our ability to learn semantic operations but also allows us to directly use a textual CLIP loss as an additional supervision when needed. We show that *pOps* can be used to learn a variety of photo-inspired operators with distinct semantic meanings, highlighting the semantic diversity and potential of our proposed approach. Code and models are available via our project page: <https://popspaper.github.io/pOps/>.

## 1 INTRODUCTION

Operators are often among the first concepts we learn in mathematics. They offer an intuitive means to describe complex concepts and equations, accompanying us from basic arithmetic operations to advanced mathematics. In the field of visual content generation, text has emerged as the de facto interface for describing and generating complex concepts. However, attaining precise control over the generated content through language is challenging, often requiring extensive prompt engineering. Drawing inspiration from the intuitiveness of operators and classical generation approaches such as

Constructive Solid Geometry [Foley 1996], we propose an operator-based generation mechanism built on top of the CLIP [Radford et al. 2021] image embedding space.

Interestingly, as observed by Ramesh *et al.* [2022], the CLIP image embedding space is already semantically meaningful, where linear operations within this subspace yield semantically meaningful embedding representations. As illustrated in Figure 2, these operations correspond to manipulations of generated images, such as compositions or the merging of concepts. However, being a vector space, users lack direct control over the exact operations performed over embeddings residing within this space. Motivated by this observation, we propose *pOps*, a general framework for training specific operators within the CLIP [Radford et al. 2021] image embedding space, with each operator reflecting a unique semantic operation.



Fig. 2. **Averaging in latent space.** Given two images we encode them to the CLIP embedding space, average their representations, and pass the result as a condition to an image diffusion model to generate an image. As shown, averaging in latent space has semantic meaning even with no training but the meaning can change unexpectedly and is not controllable.

Importantly, all *pOps* operators share the same architecture, differing only in the training data and objective. As shall be demonstrated, this unified framework allows one to compose different semantic manipulations, providing much-needed control and flexibility over the image generation process.

We represent these manipulations using the Diffusion Prior model, introduced in DALL-E 2 [Ramesh et al. 2022]. We show that the Diffusion Prior, originally trained to map text embeddings into image embeddings, can be naturally extended and fine-tuned to accommodate other conditions. In its original training scheme, the Diffusion Prior was trained to denoise image embeddings based on either text conditions or null inputs. Intuitively, the prior needed to learn not only the properties of its input conditions but also the characteristics of a broad target domain and the relation between the two. Subsequently, when fine-tuning the model over a new condition, the model can now leverage its prior understanding of the image domain, thereby focusing on relearning the condition-specific aspect of the mapping. In fact, we show that even when fine-tuning a subset of the prior model layers, the model can still operate over new input conditions. This observation also aligns with existing literature on text-to-image diffusion models, where introducing new controls such as image embeddings (IP-Adapter [Ye et al. 2023]) or spatial controls (ControlNet [Zhang and Agrawala 2023]) can be achieved with a relatively short fine-tuning performed over a pretrained model.

To illustrate the flexibility of *pOps*, we design several operators, highlighting different potential semantic applications, including:

- (1) *The Union Operator*. Given two image embeddings representing scenes with one or multiple objects, combine the objects appearing in the scenes.
- (2) *The Texturing Operator*. Given an image embedding of an object and an image embedding of a texture exemplar, paint the object with the provided texture.
- (3) *The Scene Operator*. Given an image embedding of an object and an image embedding representing a scene layout, generate an image placing the object within a semantically similar scene.
- (4) *The Instruct Operator*. Given an image embedding of an object and a single-word adjective, apply the adjective to the image embedding, altering its characteristics accordingly.
- (5) *The Composition Operator*. Given a set of object parts (e.g., articles of clothing), create a scene composing the objects together (e.g., a complete outfit).

For each operator, we independently fine-tune the Diffusion Prior model on the corresponding task to generate the desired image embedding representation. Observe that some operators (e.g., texturing and union) can be trained by defining a paired dataset of image embeddings. However, in some instances, defining a paired dataset is impractical. As such, we show how one can train operators using supervision realized by a textual CLIP loss, eliminating the need for direct image supervision.

Finally, given a set of trained *pOps* operators, we can also compose them together to form more complex semantic operations, creating a new generation paradigm. Rather than providing all conditions simultaneously and generating the output in a single shot, we can

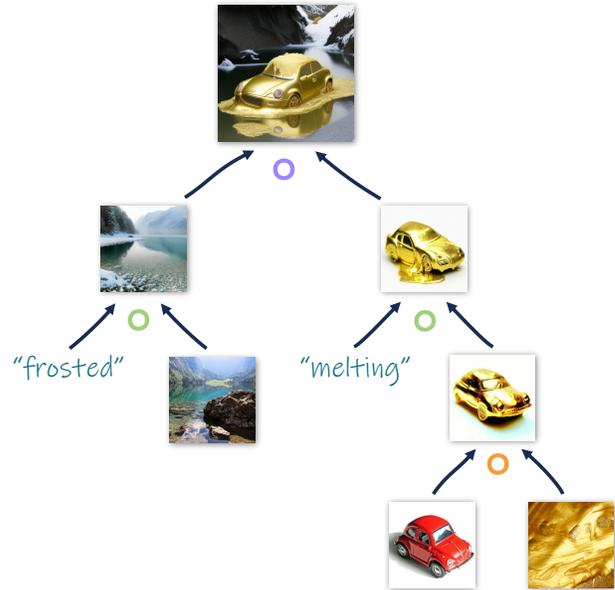


Fig. 3. *pOps* operators can be composed into *generative trees*, each node specifying a different operator applied in the CLIP image embedding space.

carefully design each element in the CLIP embedding space and compose them together into a *generative tree*. This allows users to design a more granular generation process wherein objects are first generated independently, manipulated individually, and finally merged together into a single embedding. This final embedding can then be “rendered” into a corresponding image using a pretrained image denoising network. This methodology aligns well with traditional generation processes in computer graphics, such as Constructive Solid Geometry [Foley 1996], which builds upon an iterative, tree-like modeling approach, as illustrated in Figure 3.

## 2 RELATED WORK

**Text-to-Image Generation.** Recent advancements in large-scale generative models [Po et al. 2023; Yin et al. 2024] have quickly revolutionized content creation, particularly in the domain of visual content generation. Notably, the progress in large-scale diffusion models [Balaji et al. 2023; Ding et al. 2022; Nichol et al. 2021; Ramesh et al. 2022; Rombach et al. 2022; Saharia et al. 2022; Shakhmatov et al. 2022] has resulted in unprecedented quality, diversity, fidelity. However, these models primarily rely on a free-form text prompt as guidance, often requiring extensive prompt engineering to reach the desired result [Liu and Chilton 2022; Marcus et al. 2022; Wang et al. 2022; Witteveen and Andrews 2022]. As a result, many have explored new avenues for providing users with more precise control over the generative process. This control is often realized through spatial conditions [Avrahami et al. 2023; Bar-Tal et al. 2023; Dahary et al. 2024; Huang et al. 2023; Li et al. 2023b; Voynov et al. 2023; Zhang and Agrawala 2023], including but not limited to segmentation masks, bounding boxes, and depth maps. While effective for defining structure, these methods still lack the ability to control the style and appearance of the generated image.

**Image-Conditioned Generation.** To address the limitations of text representations, some approaches aim to integrate image embeddings directly into pretrained denoising networks, most commonly through cross-attention layers. For instance, T2I-Adapter [Mou et al. 2023] controls the global style of generated images by appending image features extracted from a CLIP image encoder to the text embeddings. Similarly, Uni-ControlNet [Zhao et al. 2024] introduces an adapter tasked with projecting CLIP image embeddings to the text embedding space to achieve global control over the generated image. Most relevant to our work, IP-Adapter [Ye et al. 2023] employs a decoupled cross-attention mechanism and an Image Prompt Adapter to project image features into a pretrained text-to-image diffusion model. While all of these methods allow conditioning on image embeddings, manipulating the embeddings themselves is challenging, as they are fed into the network as-is. As a result, it remains difficult to precisely control the actual effect of this condition.

**Diffusion Prior Model.** In Ramesh et al. [2022], the authors introduce the *Diffusion Prior* model, tasked with mapping an input text embedding to a corresponding image embedding in the CLIP [Radford et al. 2021] embedding space. This image embedding is then used to condition the generative model to generate the corresponding image. This mechanism allows them to not only use existing image embeddings as a condition but also generate such inputs using a separate generative process. Originally the authors demonstrated that leveraging the Diffusion Prior leads to improved image diversity while supporting image variations, interpolation, and editing. Since then it has been shown that the prior mechanism can also be adopted for a wide range of generative tasks, including creative image generation [Richardson et al. 2023], text-to-video generation [Esser et al. 2023; Singer et al. 2023], and 3D generation [Mohammad Khalid et al. 2022; Xu et al. 2023].

**Operators and Composable Generation.** In the context of few-shot learning, Alfassy et al. [2019] demonstrate how to construct a new feature vector such that its semantic content aligns with the output of a set operation applied over a set of input vectors (e.g., intersection and union). This technique was shown to assist in few-shot discriminative settings as a form of augmentation in the feature space. In the generative domain, Composable-Diffusion [Liu et al. 2022] proposed using conjunction and negation operators to compose text prompts and better control the generation process. Concept Algebra has also been shown to be feasible in existing text-to-image models by leveraging their learned representations [Brack et al. 2024; Gandikota et al. 2023] or using a small exemplar dataset [Motamed et al. 2023; Wang et al. 2024a].

While composite generation remains an under-researched task, it has become common in the generative community to use tools such as ComfyUI and WebUI to compose different methods into a single generative scheme. In a sense, this can be viewed as a hierarchical generative process where each model serves as an operator with a dedicated task (e.g. a try-on operator [Choi et al. 2024; Xu et al. 2024]), a texturing operator ([Cheng et al. 2024]), a stylization operator [Wang et al. 2024b]). While this aligns with the inspiration behind our work, these operators are typically applied as an afterthought in the image domain, whereas we focus on manipulations in the semantic image embedding domain.

**Inspired Generation.** Human creativity has been heavily studied in the context of computer graphics, with many exploring whether computers can be used to aid the creative design process [Elhoseiny and Elfeki 2019; Esling and Devis 2020; Hertzmann 2018; Kantosalo et al. 2014; Oppenlaender 2022; Wang et al. 2024c]. At the core of the creative design process lies the ability to draw upon past knowledge to inspire the creation of novel ideas [Bonnardel and Marmèche 2005; Wilkenfeld and Ward 2001]. Crucially, this process involves associating past ideas to produce original concepts rather than simply mimicking prior work [Brown 2008; Rook and van Knippenberg 2011]. This is often achieved through the use of exemplars, drawing inspiration from their shape, color, or function.

Recently, Vinker et al. [2023] utilized a VLM to decompose a visual concept into different visual aspects, organized in a hierarchical tree structure. In doing so, they demonstrate how novel concepts and creative ideas can be discovered from a single original concept. Building on this, Lee et al. [2024] learn concept representation into disentangled language-informed axes such as category, color, and material, enabling novel concept compositions using the disentangled sub-concepts. Finally, Ng et al. [2023] extract localized sub-concepts (e.g., body parts) in an unsupervised manner that can be used to create hybrid concepts by merging the learned sub-concepts.

In this work, we focus on composing different aspects of visual concepts to inspire the generation of new visual content. This idea also draws inspiration from Constructive Solid Geometry (CSG) [Foley 1996], which combines geometric primitives via a set of boolean operators to form complex objects.

### 3 PRELIMINARIES

**Diffusion Prior.** Text-to-image diffusion models are typically trained using a conditioning vector  $c$ , which is derived from a pretrained CLIP [Radford et al. 2021] text encoder based on a user-provided text prompt  $p$ . Ramesh et al. [2022] propose a two-stage approach to the text-to-image generative process. Firstly, they train a Diffusion Prior model to map a given text embedding to a corresponding image embedding. Subsequently, the predicted image embedding is fed into a denoising diffusion probabilistic model (DDPM) [Ho et al. 2020] to generate an image.

The training process of this two-step framework resembles that of standard text-conditioned diffusion models. First, a DDPM is trained following the standard diffusion objective and aims to minimize:

$$\mathcal{L} = \mathbb{E}_{z, y, \epsilon, t} [\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2]. \quad (1)$$

Here, the denoising network  $\epsilon_\theta$  is tasked with removing the noise  $\epsilon$  added to the latent code  $z_t$  at timestep  $t$ , given the conditioning vector  $c$ , where  $c$  is now an image embedding.

Next, the Diffusion Prior model,  $P_\theta$ , is trained to predict a denoised image embedding  $e$  from a noised image embedding  $e_t$  at timestep  $t$ , given a text prompt  $y$ , by minimizing the objective given by:

$$\mathcal{L}_{prior} = \mathbb{E}_{e, y, t} [\|e - P_\theta(e_t, t, y)\|_2^2]. \quad (2)$$

In this work, we explore how the Diffusion Prior can be adapted to operate over image embeddings rather than the standard text embeddings. In doing so, we present a versatile framework capable of mapping various user inputs to their corresponding image embeddings.

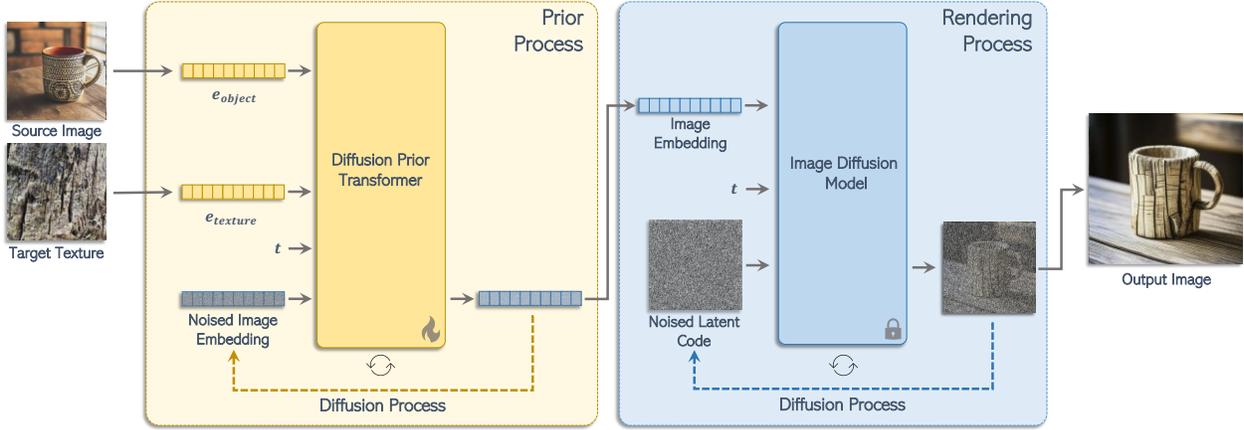


Fig. 4. **pOps Overview for the Texturing Operator.** Given an image representing our source object and an image representing our target texture, we first encode both images into the CLIP embedding space, resulting in embeddings  $e_{object}$  and  $e_{texture}$ , respectively. To train our Diffusion Prior model on the specific semantic task (shown in yellow), we perform optimization as follows. At each timestep  $t$ , we pass the two image embeddings, an encoding of  $t$ , and a noised image embedding to our Diffusion Prior model. The model is tasked with outputting a denoised image embedding that matches the target embedding  $e_{target}$ . Following training, we can pair our trained Diffusion Prior model with a pretrained, fixed image diffusion model. The learned image embedding serves as a conditioning to the diffusion model to effectively “render” the corresponding image (illustrated in blue).

## 4 THE POPS FRAMEWORK

Here, we demonstrate how  $pOps$  can be utilized to realize a variety of semantic operators. While all the  $pOps$  operators share the same architecture, they differ in terms of input conditions and corresponding training objectives.

### 4.1 Binary Image Operators

We begin with binary operators that are conditioned on two provided image embeddings and produce a single image embedding that aligns with the desired task. An overview is provided in Figure 4.

**4.1.1 Architecture and Training.** Following Ramesh *et al.* [2022], we divide the generation process into two stages. First, an image embedding is generated utilizing a dedicated transformer model. This image embedding then serves as a condition for the image diffusion model to generate the desired image. Since we work directly over image embeddings, training is required only for the prior, while the diffusion image model, acting as a “renderer”, remains fixed.

For our binary operators, the learnable task is defined using a paired dataset of input conditions,  $(I_a, I_b)$ , and a corresponding target image  $I_{target}$ , see Figure 5. These pairs represent the semantic mapping we aim to learn. As we operate in the image embedding space, we first encode all images using a pretrained CLIP image encoder [Radford *et al.* 2021],  $E_{im}(\cdot)$ , resulting in corresponding embeddings  $e_a$ ,  $e_b$ , and  $e_{target}$ . We note that the original prior model received 77 input tokens, representing the 77 text tokens extracted from the pretrained CLIP text encoder. Here, we repurpose these inputs, placing our two embeddings  $e_a$  and  $e_b$  at the start and filling the remaining entries with zero embeddings. As shall be demonstrated, reusing the original entries of the prior model allows us to adapt the number of image embeddings that we pass to the diffusion prior model to match each operator. These embeddings are followed by an encoding of the timestep  $t$  and the noised image embedding we aim to denoise. The predicted output of the prior



Fig. 5. Generated paired data for various  $pOps$  operators. During training, the images are encoded to embeddings  $e_a$ ,  $e_b$ , and  $e_{target}$ , respectively.

model is taken from the token output associated with the input noised image embedding, yellow highlighted section of Figure 4.

During training, at each optimization step, we randomly sample a timestep  $t$  and add a corresponding noise to  $e_{target}$ , resulting in the noisy image embedding  $e_{target}^t$ . We then train our prior model using the standard denoising objective:

$$\mathcal{L}_{prior} = \mathbb{E}_{e_{target}, y, t} \left[ \|e_{target} - P_{\theta}(e_{target}^t, t, e_a, e_b)\|_2^2 \right]. \quad (3)$$

Thus, our model learns to denoise  $e_{target}^t$  while taking into account the conditional embeddings  $e_a$  and  $e_b$ . During inference, we perform 25 denoising steps, starting from random noise, with an additional classifier-free guidance term where we drop the  $e_a$  and  $e_b$  inputs.

**4.1.2 Data Generation.** When trying to solve a specific image-to-image task, it is common to incorporate task-specific modules into the architecture, such as a dedicated depth estimation model applied to the input image or a background extraction model to isolate the object of interest. Instead, in  $pOps$ , we adopt a unified architecture for all our binary operators. Our model implicitly learns to manipulate the image embeddings based on the desired task. This is achieved by generating data that simulates our target task, leveraging the powerful vision and vision-language models released in recent years. Below, we outline the data generation process for the various binary operators considered in this work, with additional details and generated samples provided in Appendix A.

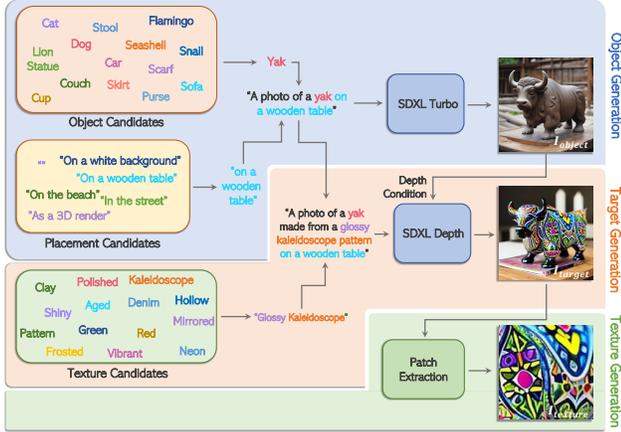


Fig. 6. **Data Generation Scheme.** An example scheme for our data generation, illustrated over our texturing operator.

**Texturing.** In the texturing operator, our input image embeddings consist of  $e_{object}$ , the embedding of the object to be textured, and  $e_{texture}$ , representing the desired texture. Our goal is to generate a target embedding  $e_{target}$ , depicting an image of  $e_{object}$  textured with  $e_{texture}$ . The data generation protocol used to create our paired texturing dataset is illustrated in Figure 6.

We begin by generating an object using SDXL-Turbo [Sauer et al. 2024]. The resulting image embedding then serves as  $e_{object}$  used during training. Next, we compile a set of attributes associated with textures and randomly sample a subset of these properties, composing them into a descriptive sentence. We then generate an image using a depth-conditioned Stable Diffusion model, conditioned on the depth of the generated object image and the composed text prompt. This process results in an image of our original object with a new texture, which we utilize to generate the embedding  $e_{target}$ . Finally, to generate  $e_{texture}$ , we automatically extract a small patch from within the target image and define it as our texture exemplar.

It is important to highlight that the texture is directly extracted from the target image. This encourages specificity, as a textual prompt can generate a range of plausible textures, whereas here, we condition the model on a *specific* texture. Furthermore, achieving a complete match between the target and object images is not necessary. For instance, there can be variations in the background between the two, as long as they remain *semantically* consistent.

**Scene.** In our scene operator, we receive two input embeddings:  $e_{object}$ , representing our object of interest, and  $e_{back}$ , denoting a target scene background for placing the object. As in texturing, we initially generate an image of our object using SDXL-Turbo, which corresponds to  $e_{target}$ . Next, we employ a background removal model [BRIA 2024] to isolate our object from the generated image. The segmented object is then positioned either on a white background or within a newly generated background, which is encoded into the  $e_{object}$  embedding. Lastly, we utilize a Stable Diffusion inpainting model to produce an image containing only the original background, which we encode to  $e_{back}$ . In essence, during the data generation phase, we decompose the target into separate representations of its object and background. Through this process, *pOps* can learn how to effectively compose the two elements back together.

**Union.** In our union operator, we receive two image embeddings representing two objects, denoted as  $e_a$  and  $e_b$ , with the aim of generating an image embedding that plausibly incorporates both objects. To construct the union dataset, we build on the intuition that separating objects from existing scenes is typically easier than integrating them together. Therefore, we first construct a dataset of images containing pairs of objects by randomly selecting two object classes and generating an image containing both objects using SDXL-Turbo (e.g., “a cat and a banana”). This resulting image is then encoded to define  $e_{target}$ . Next, we employ a grounded detection method, OWLv2 [Minderer et al. 2024], to extract each object of interest as an individual crop, generating  $e_a$  and  $e_b$ , respectively. The *pOps* operator is then tasked with composing these part embeddings back into a single image combining both parts.

## 4.2 Multi-Image Compositions

While binary operators cover a wide range of tasks and can be combined in a tree-like structure to execute more complex operations, some operators can benefit from considering all inputs simultaneously. To illustrate this, we explore a specific composition operator that takes a set of embeddings, each representing a distinct clothing item, and combines them into a single representation of a person wearing those clothes. To train such an operator, we extend the input sequence to accommodate the set of clothing items, setting a fixed input index for each clothing type. This again leverages the original design of the prior, which was tailored to process a sequence of 77 input text tokens. For training, we utilize the ATR dataset [Liang et al. 2015], developed for human parsing. We encode the given complete image as our target embedding  $e_{target}$  and decompose the clothing items using the segmentation masks annotated in the dataset to form our input sequence. The training scheme itself is identical to the binary operators, utilizing Equation (4) to train the prior model on our composition task.

## 4.3 The Instruct Operator

All the operators discussed so far have assumed a paired dataset with a well-defined target embedding. However, operating in the CLIP embedding space presents interesting opportunities to easily apply additional losses within this space. In particular, we explore a binary operator that takes as input a CLIP **image** embedding of an object, denoted as  $e_{object}$ , and a CLIP **text** embedding of a target adjective, labeled as  $e_{instruct}$  (e.g., “spiky”, “hairy”, “melting”). With these inputs, the prior model is tasked with generating an embedding  $e_{object}$  corresponding to an image portraying the adjective described in  $e_{instruct}$ .

Since both the image embedding and text embedding reside in a shared CLIP space of the same dimensionality, we can easily feed both into our transformer. To train our task, we introduce an additional loss objective that evaluates the CLIP similarity between the generated image and the embedding  $e_{text}$  of the prompt combining the target adjective and object class (e.g., “a spiky dog”). Formally, our new loss objective is given by:

$$\mathcal{L} = \mathcal{L}_{prior} + \lambda(e_{text}, P_{\theta}(e_c^t, e_{object}, e_{instruct})), \quad (4)$$

where  $P_{\theta}(e_c^t, e_{object}, e_{instruct})$  is the generated embedding.

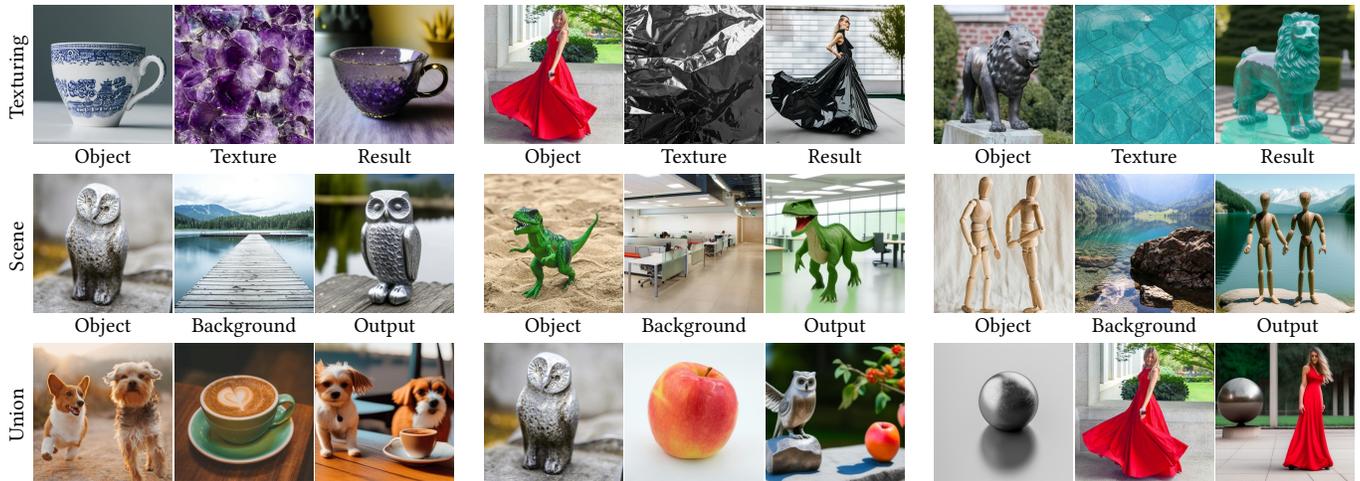


Fig. 7. Results obtained with our binary *pOps* operators. Notice that while images are visualized, all operations are applied within the embedding space.

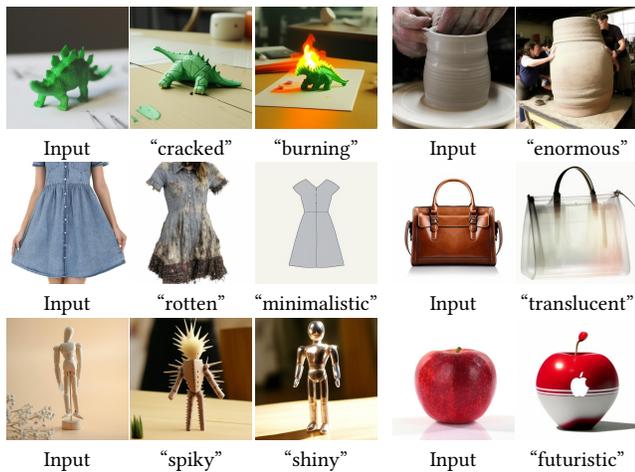


Fig. 8. Instruct operator results obtained by *pOps*.

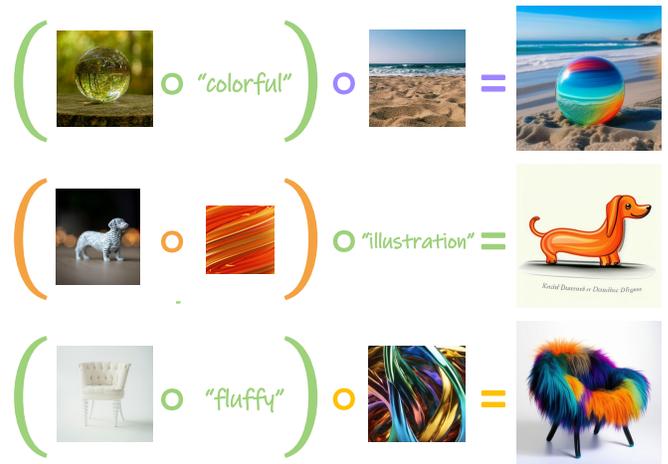


Fig. 9. Multi-operator compositions obtained by our *pOps* method.

## 5 EXPERIMENTS

We now turn to validate the effectiveness of *pOps* through a comprehensive set of evaluations. Additional details, along with a large gallery of results, are available in Appendices A, C and D.

**Operator Results.** Results for our binary operators are provided in Figure 7, where each operator effectively realizes a specific and consistent semantic operation. Given that we operate within the CLIP embedding space, the operators focus on preserving the semantic nature of the inputs while being agnostic to the structure or placement of the objects. Next, we present results for our instruct operator in Figure 8. Given a single descriptive word, our operator successfully generates a plausible output incorporating both the adjective and input object. As shown in Figure 9, our operators can also be combined into generative equations representing more complex semantic operations. These operations are applied directly in the image embedding space, where only the final embedding is “rendered” into a corresponding image. Finally, Figure 10 shows a



Fig. 10. Multi-image compose operator results obtained by *pOps*.

multi-input example where *pOps* was trained to take a sequence of embeddings corresponding to articles of clothing and output an embedding that represents the complete outfit. Additional results for all operators are available in Appendix D.



Fig. 11. Qualitative comparison of  $pOps$  to latent averaging.

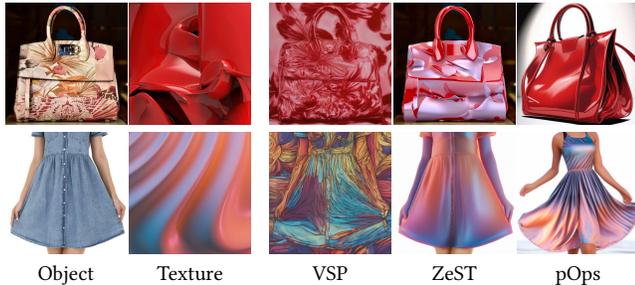


Fig. 12. Qualitative comparison for the  $pOps$  texturing operator.

**Qualitative Comparisons.** Next, we evaluate our union and scene operators in comparison to latent space averaging. As can be seen in Figure 11,  $pOps$  applies a consistent operation to the provided inputs, whereas averaging yields outputs with varying semantic meanings. This observation aligns with our expectations that the CLIP embedding space is well-suited for semantic operations but is inconsistent when used naïvely. We proceed to evaluate our texturing and instruct operators by comparing them to relevant literature. In Figure 12, we compare our texturing operator to Visual Style Prompting [Jeong et al. 2024] and ZeST [Cheng et al. 2024].

Similarly, in Figure 13, we compare our instruct operator to InstructPix2Pix [Brooks et al. 2023] and IP-Adapter [Ye et al. 2023]. Note that  $pOps$  has seen the instructions during training, but without direct supervision that was used in InstructPix2Pix. Comparisons to additional baselines can be found in Appendix C.

**Quantitative Comparisons.** We conduct two forms of quantitative evaluation to validate the effectiveness of our approach. In Table 1, we utilize image and text similarity metrics to compare our instruct operator to InstructPix2Pix and IP-Adapter. One can see that our method attains higher image similarity than IP-Adapter with a scale of 0.1 while still retaining high text similarity values. Next, we perform a user study for the instruct and texturing tasks alongside their alternatives. The results in Table 2 demonstrate that  $pOps$  compares favorably to the recent state-of-the-art in both tasks.

**Analysis.** As discussed, the image generation process in  $pOps$  is independent of the trainable operator itself. Therefore, we have the flexibility to employ any compatible image generation model that can be conditioned on our CLIP image embeddings. While our model of choice was Kandinsky 2 [Shakhmatov et al. 2022], in Figure 14 we show that our method is also compatible with IP-Adapter without



Fig. 13. Qualitative comparison for the instruct operator to existing approaches: InstructPix2Pix [Brooks et al. 2023] & IP-Adapter [Ye et al. 2023].

Table 1. Quantitative Comparison for the Instruct Operator. Image similarity is computed with DreamSim [Fu et al. 2023] and text similarity with CLIP ViT-L/14. Results are averaged across 52 objects and 65 adjectives.

Method	Image Similarity $\uparrow$	Text Similarity $\uparrow$	BERT Similarity $\uparrow$
InstructPix2Pix	0.455	0.237	0.424
IP-Adapter (0.5)	0.826	0.211	0.544
IP-Adapter (0.1)	0.584	0.219	0.531
$pOps$	0.6607	0.236	0.437

Table 2. User study results for the instruct and texturing operators.

Metric	Instruct			
	InstructP2P	IP-Adapter	IP-Adapter (0.1)	$pOps$
Percent Preferred $\uparrow$	23.81%	3.18%	12.70%	<b>60.31%</b>
Average Rating $\uparrow$	1.65	1.95	2.75	<b>3.49</b>
Metric	Texturing			
	IP-Adapter	VSP	ZeST	$pOps$
Percent Preferred $\uparrow$	3.57%	1.79%	37.50%	<b>57.14%</b>
Average Rating $\uparrow$	1.45	2.21	3.66	<b>3.98</b>

requiring any modification or tuning. This compatibility enables us to leverage a diverse range of models supported by IP-Adapter, including a depth-conditioned ControlNet.

Finally, since  $pOps$  employs a diffusion model to generate image embeddings, we can sample different seeds for the same input conditions. Interestingly, in Figure 15, we demonstrate that when providing only a single input to the texturing operator, the model can sample diverse and plausible results based on the given input. Additional examples for both analyses are provided in Appendix D.



Fig. 14. **Different Renderers.**  $pOps$  outputs can be directly fed to either Kandinsky or IP-Adapter and incorporated alongside spatial conditions

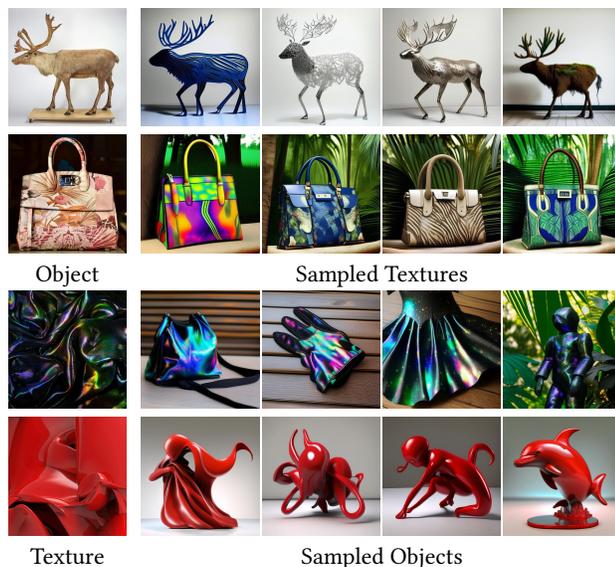


Fig. 15. **Sampling from missing inputs.** Given only an object or a texture, the  $pOps$  texturing operator can successfully sample diverse textured objects.

## 6 LIMITATIONS

While our experiments highlight the potential of  $pOps$  for semantic control, it is important to also discuss the limitations of our approach. First, there are inherent limitations when operating within the CLIP domain. As previously discussed in Ramesh *et al.* [2022], the semantic embedding fails to preserve some visual attributes. In Figure 16 we visualize these limitations by viewing direct reconstructions of images when passing them through the CLIP embedding space. Although the embedding space effectively encodes the objects semantically, it struggles with encoding their distinct visual appearance compared to optimization-based personalization methods. As shown, CLIP also struggles with binding two different visual attributes to two distinct objects. This was most evident in our results for the union operation where the “rendered” result may leak colors between the two objects, struggling with maintaining the distinct appearance of each one.

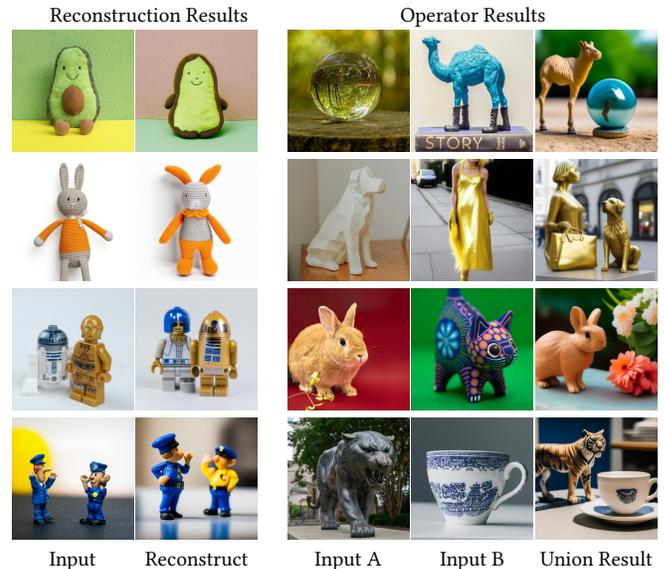


Fig. 16. **Limitations of  $pOps$ .** On the left we show reconstructions achieved by directly embedding an image into CLIP and reconstructing it with Kandinsky2, highlighting the limitations of the embedding space. On the right, we show failure cases for our union operator, where attribute leakage is visible or where the operator struggles with preserving both objects.

Additionally,  $pOps$  tunes each operator independently, where it might be more beneficial to train a single diffusion model capable of realizing all of our different operators together or alternatively do only a low-rank adaptation [Hu *et al.* 2021] when training an operator. Finally, all  $pOps$  operators were trained on a single GPU for a few days. This leads us to believe that further computational scaling could potentially improve performance even within the limitations of the CLIP space and current architecture.

## 7 CONCLUSIONS

In this work, we have introduced  $pOps$ , a framework designed for training semantic operations directly on CLIP image embeddings.  $pOps$  offers a new take on image generation, providing users with specific forms of semantic control over image embeddings that can then be joined together to form the desired concept. Our method builds upon both generated datasets that represent the task at hand and can also be supervised directly using a CLIP-based objective. We believe that  $pOps$  opens up new possibilities for training a wide variety of operators within the CLIP space and other semantic spaces. These new operators can then be composed with one another to create even more creative possibilities along the generation process.

## ACKNOWLEDGMENTS

We would like to thank Rinon Gal, Yael Vinker, and Or Patashnik for their discussions and valuable input which helped improve this work. We would also like to thank Andrey Voynov for his early feedback on this work. This work was supported by the Israel Science Foundation under Grant No. 2366/16 and Grant No. 2492/20.

## REFERENCES

- Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. 2023a. Cross-image attention for zero-shot appearance transfer. *arXiv preprint arXiv:2311.03335* (2023).
- Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. 2023b. A Neural Space-Time Representation for Text-to-Image Personalization. arXiv:2305.15391 [cs.CV]
- Amit Alfassy, Leonid Karlinsky, Amit Aides, Joseph Shtok, Sivan Harary, Rogerio Feris, Raja Giryes, and Alex M Bronstein. 2019. Laso: Label-set operations networks for multi-label few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6548–6557.
- Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. 2023. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18370–18380.
- Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. 2023. eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. arXiv:2211.01324 [cs.CV]
- Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. 2023. Multidiffusion: Fusing diffusion paths for controlled image generation. (2023).
- Nathalie Bonnardel and Evelyne Marmèche. 2005. Towards supporting evocation processes in creative design: A cognitive approach. *International journal of human-computer studies* 63, 4-5 (2005), 422–435.
- Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. 2024. SEGA: Instructing text-to-image models using semantic guidance. *Advances in Neural Information Processing Systems* 36 (2024).
- BRIA. 2024. *BRIA Background Removal v1.4*. <https://huggingface.com/briai/RMBG-1.4>
- Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *CVPR*.
- David C Brown. 2008. Guiding computational design creativity research. *Studying Design Creativity*, Springer (2008).
- Ta-Ying Cheng, Prafull Sharma, Andrew Markham, Niki Trigoni, and Varun Jampani. 2024. ZeST: Zero-Shot Material Transfer from a Single Image. *arXiv preprint arXiv:2404.06425* (2024).
- Yisul Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. 2024. Improving Diffusion Models for Virtual Try-on. *arXiv preprint arXiv:2403.05139* (2024).
- Omer Dahary, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. 2024. Be Yourself: Bounded Attention for Multi-Subject Text-to-Image Generation. *arXiv preprint arXiv:2403.16990* (2024).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. 2022. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems* 35 (2022), 16890–16902.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=YicbFdNTTy>
- Mohamed Elhoseiny and Mohamed Elfeki. 2019. Creativity inspired zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5784–5793.
- Philippe Esling and Ninon Devis. 2020. Creativity in the era of artificial intelligence. *arXiv preprint arXiv:2008.05959* (2020).
- Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. 2023. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011* (2023).
- James D Foley. 1996. 12.7 Constructive solid geometry. , 533–558 pages.
- Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. 2023. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. arXiv:2306.09344 [cs.CV]
- Rohit Gandikota, Joanna Materzynska, Tingrui Zhou, Antonio Torralba, and David Bau. 2023. Concept sliders: Lora adaptors for precise control in diffusion models. *arXiv preprint arXiv:2311.12092* (2023).
- Aaron Hertzmann. 2018. Can computers create art?. In *Arts*, Vol. 7. MDPI, 18.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718* (2021).
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).
- Lianghua Huang, Di Chen, Yu Liu, Shen Yujun, Deli Zhao, and Zhou Jingren. 2023. Composer: Creative and Controllable Image Synthesis with Composable Conditions. (2023).
- Jaeseok Jeong, Junho Kim, Yunjey Choi, Gayoung Lee, and Youngjung Uh. 2024. Visual Style Prompting with Swapping Self-Attention. arXiv:2402.12974 [cs.CV]
- Anna Kantosalo, Jukka M Toivanen, Ping Xiao, and Hannu Toivonen. 2014. From Isolation to Involvement: Adapting Machine Creativity Software to Support Human-Computer Co-Creation. In *ICCC*. 1–7.
- Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV* (2020).
- Sharon Lee, Yunzhi Zhang, Shangzhe Wu, and Jiajun Wu. 2024. Language-Informed Visual Concept Learning. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=juuyW8B8ig>
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.
- Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023b. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22511–22521.
- Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan. 2015. Deep human parsing with active template regression. *IEEE transactions on pattern analysis and machine intelligence* 37, 12 (2015), 2402–2414.
- Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. 2022. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*. Springer, 423–439.
- Vivian Liu and Lydia B Chilton. 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *CHI Conference on Human Factors in Computing Systems*. 1–23.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=Bkg6RiCqY7>
- Gary Marcus, Ernest Davis, and Scott Aaronson. 2022. A very preliminary analysis of DALL-E 2. *arXiv preprint arXiv:2204.13807* (2022).
- Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. 2024. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems* 36 (2024).
- Matthias Minderer, Alexey A. Gritsenko, and Neil Houlsby. 2023. Scaling Open-Vocabulary Object Detection. In *Thirty-seventh Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=mQPNcBwjGc>
- Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. 2022. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 conference papers*. 1–8.
- Saman Motamed, Danda Pani Paudel, and Luc Van Gool. 2023. Lego: Learning to Disentangle and Invert Concepts Beyond Object Appearance in Text-to-Image Diffusion Models. *arXiv preprint arXiv:2311.13833* (2023).
- Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaoju Qie. 2023. T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. arXiv:2302.08453 [cs.CV]
- Kam Woh Ng, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. 2023. DreamCreature: Crafting Photorealistic Virtual Creatures from Imagination. arXiv:2311.15477 [cs.CV]
- Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).
- Jonas Oppenlaender. 2022. The Creativity of Text-to-Image Generation. In *Proceedings of the 25th International Academic Mindtrek Conference (Academic Mindtrek 2022)*. ACM. <https://doi.org/10.1145/3569219.3569352>
- Ryan Po, Wang Yifan, Vladislav Golyanik, Kfir Aberman, Jonathan T. Barron, Amit H. Bermano, Eric Ryan Chan, Tali Dekel, Aleksander Holynski, Angjoo Kanazawa, C. Karen Liu, Lingjie Liu, Ben Mildenhall, Matthias Nießner, Björn Ommer, Christian Theobalt, Peter Wonka, and Gordon Wetzstein. 2023. State of the Art on Diffusion Models for Visual Computing. arXiv:2310.07204 [cs.AI]
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).
- Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. <https://arxiv.org/abs/1908.10084>

- Elad Richardson, Kfir Goldberg, Yuval Alaluf, and Daniel Cohen-Or. 2023. Conceptlab: Creative generation using diffusion prior constraints. *arXiv preprint arXiv:2308.02669* (2023).
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. , 10684–10695 pages.
- Laurens Rook and Daan van Knippenberg. 2011. Creativity and imitation: Effects of regulatory focus and creative exemplar quality. *Creativity Research Journal* 23, 4 (2011), 346–356.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494.
- Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. 2024. Fast High-Resolution Image Synthesis with Latent Adversarial Diffusion Distillation. *arXiv preprint arXiv:2403.12015* (2024).
- Arseniy Shakhmatov, Anton Razzhigaev, Aleksandr Nikolich, Vladimir Arkhipkin, Igor Pavlov, Andrey Kuznetsov, and Denis Dimitrov. 2022. Kandinsky 2. <https://github.com/ai-forever/Kandinsky-2>.
- Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. 2023. Make-A-Video: Text-to-Video Generation without Text-Video Data. In *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=nJfyIDvgzLq>
- Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. 2023. Concept decomposition for visual exploration and inspiration. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–13.
- Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. 2023. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.
- Haofan Wang, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. 2024b. InstantStyle: Free Lunch towards Style-Preserving in Text-to-Image Generation. *arXiv preprint arXiv:2404.02733* (2024).
- Haonan Wang, James Zou, Michael Mozer, Linjun Zhang, Anirudh Goyal, Alex Lamb, Zhun Deng, Michael Qizhe Xie, Hannah Brown, and Kenji Kawaguchi. 2024c. Can AI Be as Creative as Humans? *arXiv preprint arXiv:2401.01623* (2024).
- Zihao Wang, Lin Gui, Jeffrey Negrea, and Victor Veitch. 2024a. Concept Algebra for (Score-Based) Text-Controlled Generative Models. *Advances in Neural Information Processing Systems* 36 (2024).
- Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2022. DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models. *arXiv preprint arXiv:2210.14896* (2022).
- Merryl J Wilkenfeld and Thomas B Ward. 2001. Similarity and emergence in conceptual combination. *Journal of Memory and Language* 45, 1 (2001), 21–38.
- Sam Witteveen and Martin Andrews. 2022. Investigating Prompt Engineering in Diffusion Models. *arXiv preprint arXiv:2211.15462* (2022).
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. <https://www.aclweb.org/anthology/2020.emnlp-demos.6>
- Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. 2023. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20908–20918.
- Yuhao Xu, Tao Gu, Weifeng Chen, and Chengcai Chen. 2024. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. *arXiv preprint arXiv:2403.01779* (2024).
- Hu Ye, Jun Zhang, Sibio Liu, Xiao Han, and Wei Yang. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. (2023).
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A Survey on Multimodal Large Language Models. [arXiv:2306.13549 \[cs.CV\]](https://arxiv.org/abs/2306.13549)
- Lvmin Zhang and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543* (2023).
- Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. 2024. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems* 36 (2024).

# Appendix

## A ADDITIONAL DETAILS

### A.1 Implementation Details

**Models and Architectures.** In this work, we use the CLIP ViT-bigG-14-laion2B-39B-b160k model [Dosovitskiy et al. 2021; Radford et al. 2021] for our embedding space, implemented using the Transformers library [Wolf et al. 2020]. The architecture of our Diffusion Prior model follows the same architecture as used in Kandinsky 2 [Shakhmatov et al. 2022]. For our diffusion models, we show results over both the Kandinsky 2.2 model [Shakhmatov et al. 2022] and IP-Adapter [Ye et al. 2023], both of which support this specific CLIP model.

**Training Scheme.** We train all models using a batch size of 1 over a single GPU. The models are trained using the AdamW optimizer [Loshchilov and Hutter 2019] with a constant learning rate of  $1e-5$ . Each operator is trained for approximately 500,000 training steps when trained from scratch. However, we found, empirically, that fine-tuning the model from an existing operator rather than the original Diffusion Prior model speeds up convergence. Unless otherwise noted, we train all the layers of the Diffusion Prior model.

### A.2 Data Generation

In the main paper, we discuss the process used for generating data for each operator. Below, we provide additional details. Samples of the generated data are illustrated in Figure 17. Unless otherwise noted, for each operator, we generate approximately 50,000 samples.

**Texturing.** The data generation scheme for our texturing operator is illustrated in Figure 5 of the main paper. We consider 290 object candidates across various categories such as geometric objects, animals, statues, and other miscellaneous common objects. We additionally consider 24 different object placement candidates and 310 texture attributes. For generating the object images, we use SDXL-Turbo [Sauer et al. 2024] and use prompts of the form “A photo of a  $\langle \text{object} \rangle$   $\langle \text{placement} \rangle$ ”.

To generate the target image  $I_{\text{target}}$ , we then sample between one and five texture attributes and generate an image using a depth-conditioned Stable Diffusion 2.0 [Rombach et al. 2022] model using prompts of the form “A photo of a  $\langle \text{object} \rangle$  made from  $\langle \text{texture 1}, \text{texture 2}, \dots \rangle$   $\langle \text{placement} \rangle$ .” The generation process is conditioned on the depth map extracted from  $I_{\text{object}}$ .

Finally, we are left to extract the patch representing our input texture image  $I_{\text{texture}}$ . To this end, we first detect the object in the generated image using an OWLv2 [Minderer et al. 2023] model with the prompt “A  $\langle \text{object} \rangle$ ”. We then select a small patch from within the output bounding box and use this as our texture image.

**Scene.** For our scene operator, as noted in the main paper,  $I_{\text{object}}$  is created either by pasting the segmented object either on a white background or a newly generated background. For the newly generated background, we compose a set of 208 possible backgrounds such as “On the beach”, “On the farm”, “In the castle”, etc. For our inpainting model, used to create  $I_{\text{back}}$ , we employ the SD-XL inpainting 0.1 model using the mask extracted from our object.

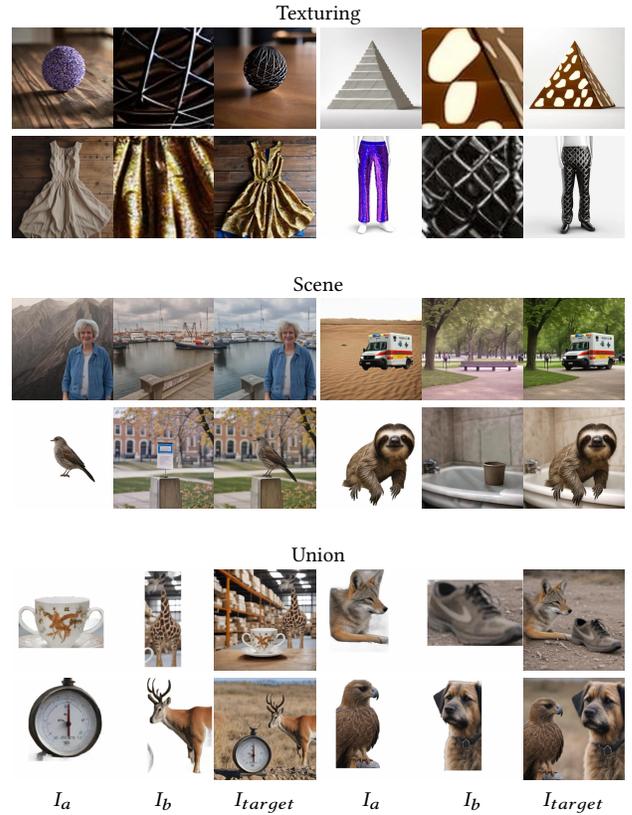


Fig. 17. Generated paired data for various  $pOps$  operators. During training, the images are encoded to embeddings  $e_a$ ,  $e_b$ , and  $e_{\text{target}}$ , respectively.

**Union.** For generating our union dataset, we consider 20,000 different objects, taken from the raw classes list from Open Images [Kuznetsova et al. 2020].

**Instruct.** Here, we sample our images from a set of 20,000 possible classes, as above, and a list of 60 possible adjectives.

**Composition.** As noted in the main paper, for our composition operator, we use the ATR dataset [Liang et al. 2015] for training. In total, we use 17,000 images for training, comprising 12 different clothing categories.

## B EVALUATION SETUP

### B.1 Baseline Methods

**Texturing.** For evaluating our texturing operator, we consider four alternative methods: (1) Cross-Image Attention [Alaluf et al. 2023a], (2) IP-Adapter [Ye et al. 2023], (3) Visual Style Prompting [Jeong et al. 2024], and (4) ZeST [Cheng et al. 2024].

For all methods, we use their official implementation and default hyperparameters. For IP-Adapter [Ye et al. 2023], we consider IP-Adapter trained over Stable Diffusion 1.5 [Rombach et al. 2022] which uses OpenCLIP-ViT-H-14 for extracting the conditioning image embeddings.

**Instruct.** For the instruct operator, we consider three approaches: (1) IP-Adapter [Ye et al. 2023], (2) InstructPix2Pix [Brooks et al. 2023], and (3) NeTI [Alaluf et al. 2023b].

For IP-Adapter, we consider two variants. First, we use the IP-Adapter Plus variant trained over Stable Diffusion 1.5 using a scale of 0.5, where we pass the adjective as the guiding text prompt. However, we attained better results when using the more recent IP-Adapter for SDXL 1.0 which is conditioned on image embeddings extracted from OpenCLIP-ViT-H-14 (ip-adapter-plus\_sd-xl\_vit-h). We found that to achieve meaningful semantic modifications, a low scale factor of 0.1 was needed. However, when doing so, the resulting images generated by IP-Adapter no longer resembled the original images. As such, we captioned the original images using BLIP-2 [Li et al. 2023a] and passed the image caption along with the desired adjective to IP-Adapter as the guiding text prompt. We found that this allowed for better alignment with the adjective (thanks to the low scale) while better preserving the original image (thanks to the image caption).

Finally, we compare *pOps* to NeTI, an optimization-based personalization method (see Figure 23). We follow the default hyperparameters and train a new concept using the image of the object. The best results were achieved when training for 250 optimization steps, as additional training led to overfitting the original image. At inference, we generated images using prompts of the form “A photo of a <adjective>  $S_*$ ”. When needed, we manually modified the prompts to ensure that they were grammatically correct.

## B.2 Quantitative Evaluations

Below we provide details regarding the evaluation data and protocol reported in the main paper.

**Texturing.** To quantitatively evaluate performance on the texturing task, we consider 52 images of objects spanning various categories including animals, statues, food items, accessories, and more. For each object, we paint the object using 16 different texture patches, resulting in 832 object-texture combinations. For each of the considered methods, we utilized three different random seeds, which gave 2, 496 total results.

As no standard metric exists for evaluating the quality of the texturing, we perform a perceptual user study. We consider two types of questions: (1) top preference and (2) rating. More specifically, users were first shown the results of the four methods side-by-side and asked to choose the result they most preferred while taking into account both how the original object was preserved and how the target texture was applied. Next, users were asked to rate the result of each method on a scale of 1 to 5, with 5 being the best, on how well the original object was preserved and the texture was applied to it. Each user was shown 7 questions for each of the two types.

**Instruct.** To evaluate our instruct operator, we similarly construct an evaluation set. Here, we consider the same 52 objects as above and construct a set of 65 adjectives. We then modify each of the 52 objects with each adjective, resulting in 3, 380 combinations. As above, each method is applied using three different seeds, resulting in 10, 140 generated images.

For our evaluation metric, we first consider the standard CLIP-Score [Hessel et al. 2021] and measure CLIP-space similarities. Specifically, we first compute the image similarity between the generated images and the original image. Next, we calculate the CLIP-space similarity between the embeddings of the generated images and the embedding of text prompts of the form “A <adjective> photo”. Finally, we consider an additional text-based similarity metric. Here, we first manually create a short caption of the target object (e.g., “A lion statue”, “A dress”). We then caption the generated images using BLIP-2 [Li et al. 2023a]. We then compute a sentence similarity measure [Devlin et al. 2018; Reimers and Gurevych 2019], computing the average cosine similarity between sentence embeddings extracted from the generated caption and captions of the form “A photo of a <adjective> <caption>.” This metric was designed to better capture the ability of the methods to integrate the desired adjective while preserving the original object class.

## C ADDITIONAL COMPARISONS

We provide additional qualitative comparisons, as follows:

- (1) First, in Figures 18 to 20, we provide additional comparisons over our binary operators (union, scene, and texturing), comparing our *pOps* results with those obtained from a simple latent averaging within the CLIP embedding space.
- (2) In Figure 21, we provide additional qualitative comparisons to alternative texturing approaches: Cross-Image Attention [Alaluf et al. 2023a], IP-Adapter [Ye et al. 2023], Visual Style Prompting [Jeong et al. 2024], and ZeST [Cheng et al. 2024].
- (3) In Figure 22, we show additional qualitative comparisons over our instruct operator, comparing *pOps* to two alternative approaches: InstructPix2Pix [Brooks et al. 2023] and IP-Adapter [Ye et al. 2023].
- (4) Finally, in Figure 23, we compare our instruct operator to an additional optimization-based personalization approach, NeTI [Alaluf et al. 2023b].

## D ADDITIONAL RESULTS

Finally, in the below Figures, we provide additional results:

- (1) In Figure 24 and Figure 25, we show additional results obtained by our texturing Diffusion Prior model when using null inputs for the object input and texture input, respectively.
- (2) In Figures 26 to 28, we provide additional texturing results.
- (3) In Figure 29, we provide additional union results.
- (4) In Figures 30 to 32, we show additional scene operator results.
- (5) In Figure 33, we provide additional instruct operator results.
- (6) In Figures 34 and 35, we show additional multi-image clothing composition results obtained with *pOps*.
- (7) In Figure 36, we show results obtained with both Kandinsky [Shakhmatov et al. 2022] and IP-Adapter [Ye et al. 2023] as renderers as well as results obtained with IP-Adapter alongside ControlNet [Zhang and Agrawala 2023] with depth-conditioning.
- (8) Finally, in Figures 37 to 39, we provide examples of operator compositions, combining our scene, instruct, and texturing *pOps* operators.

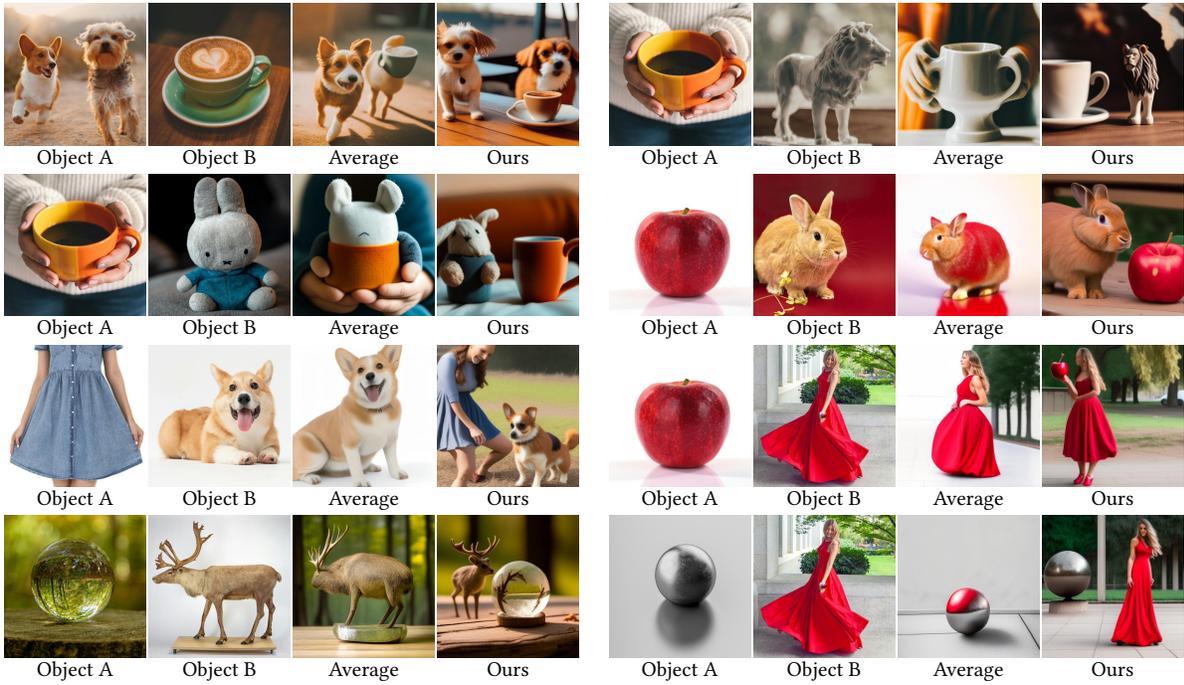


Fig. 18. Qualitative comparison of our *pOps* union operator compared to results obtained by averaging over the CLIP image embeddings.

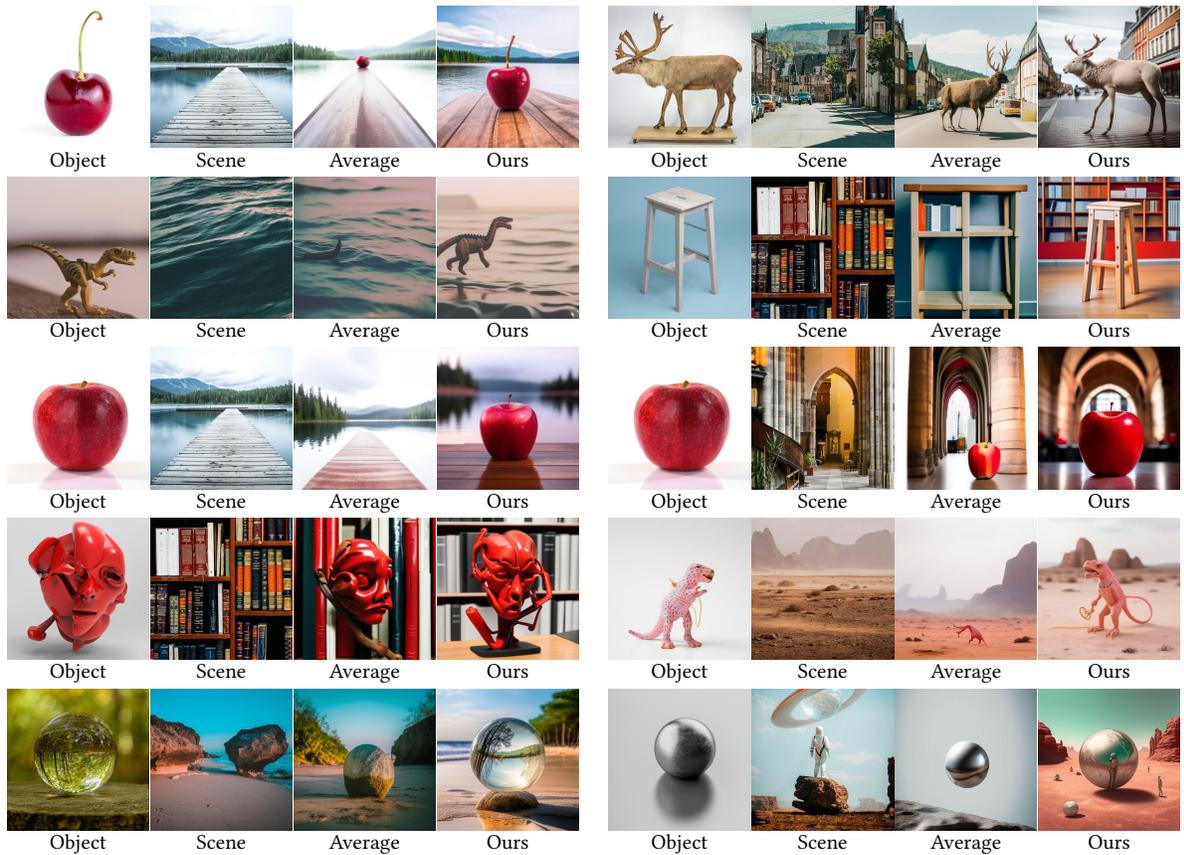


Fig. 19. Qualitative comparison of our *pOps* scene operator compared to results obtained by averaging over the CLIP image embeddings.

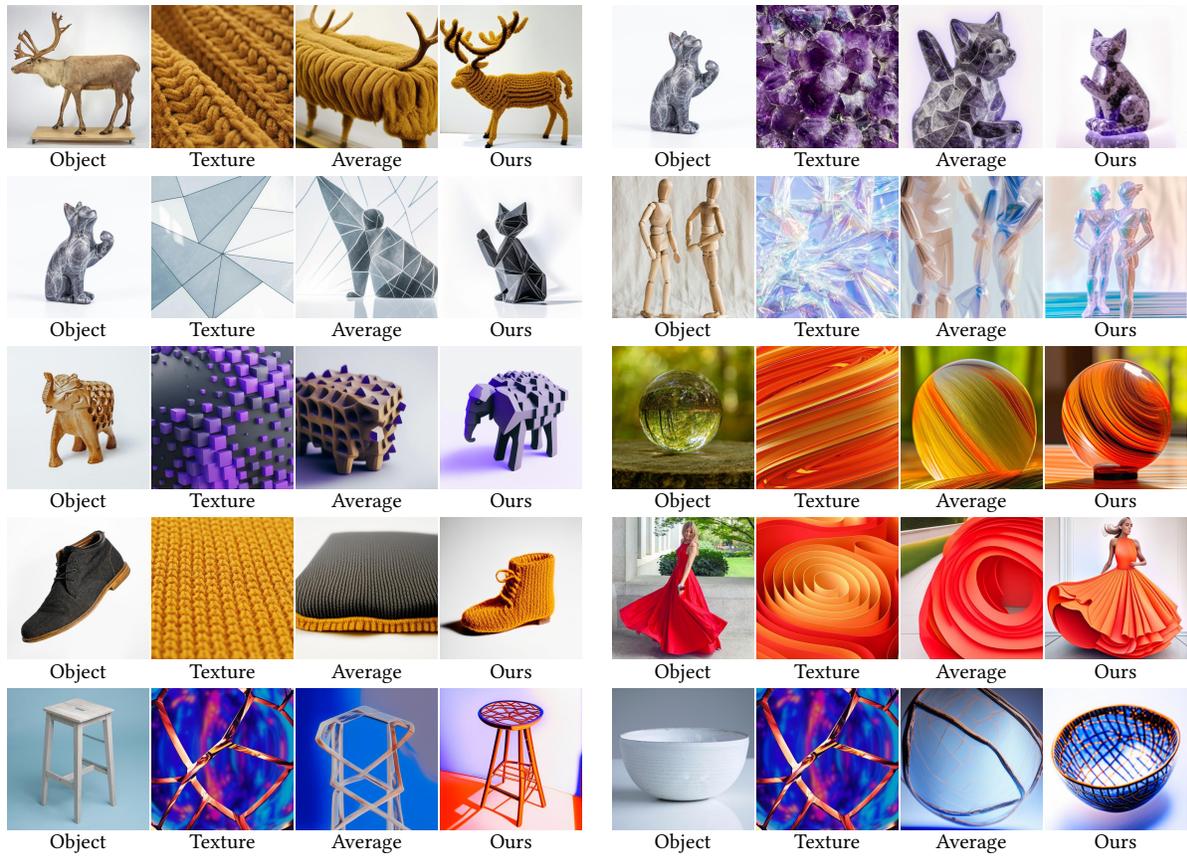


Fig. 20. Qualitative comparison of our *pOps* texturing operator compared to results obtained by averaging over the CLIP image embeddings.

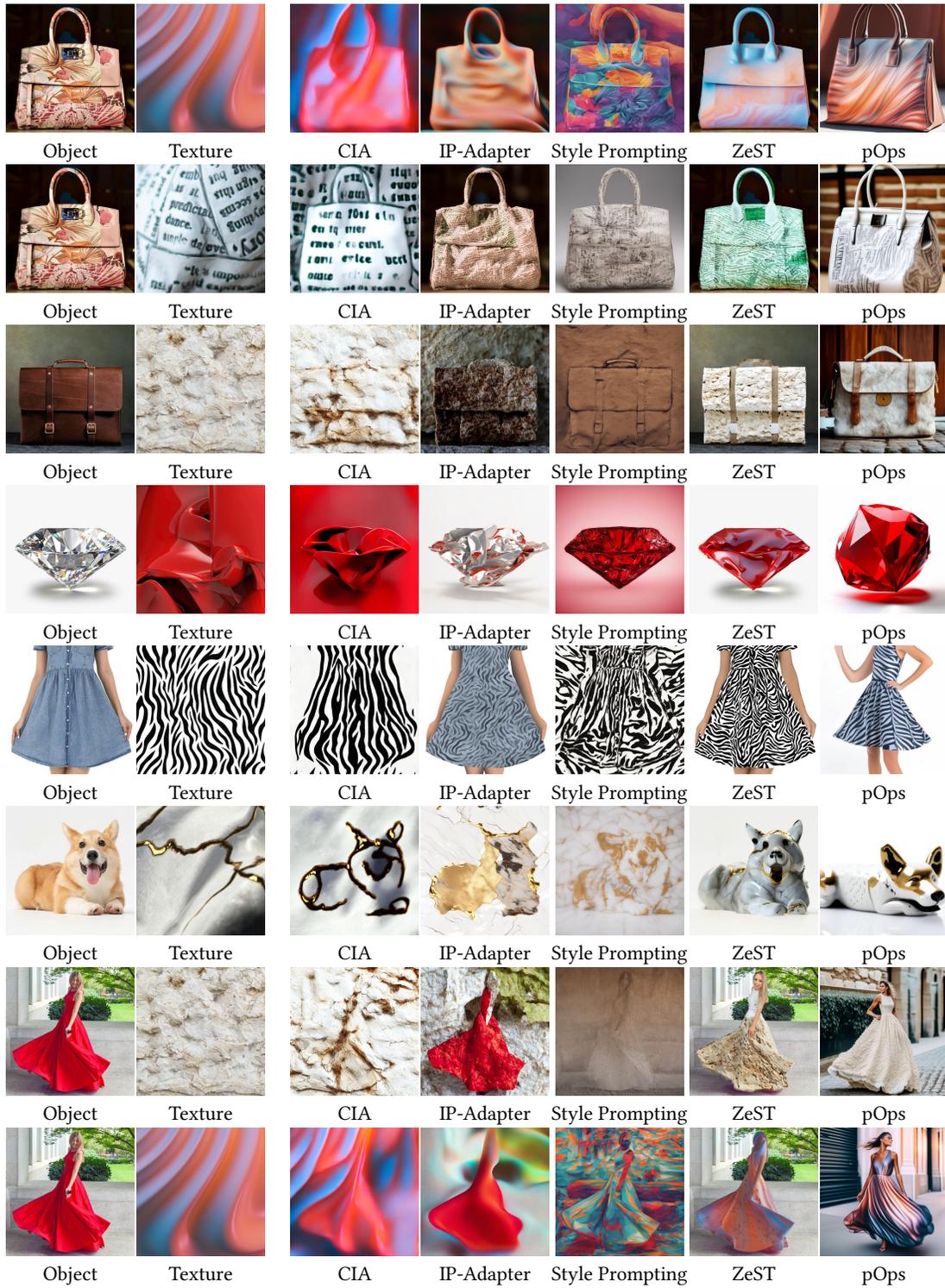


Fig. 21. Additional qualitative comparison for the pOps texturing operator to alternative texturing approaches: Cross-Image Attention [Alaluf et al. 2023a], IP-Adapter [Ye et al. 2023], Visual Style Prompting [Jeong et al. 2024], and ZeST [Cheng et al. 2024].

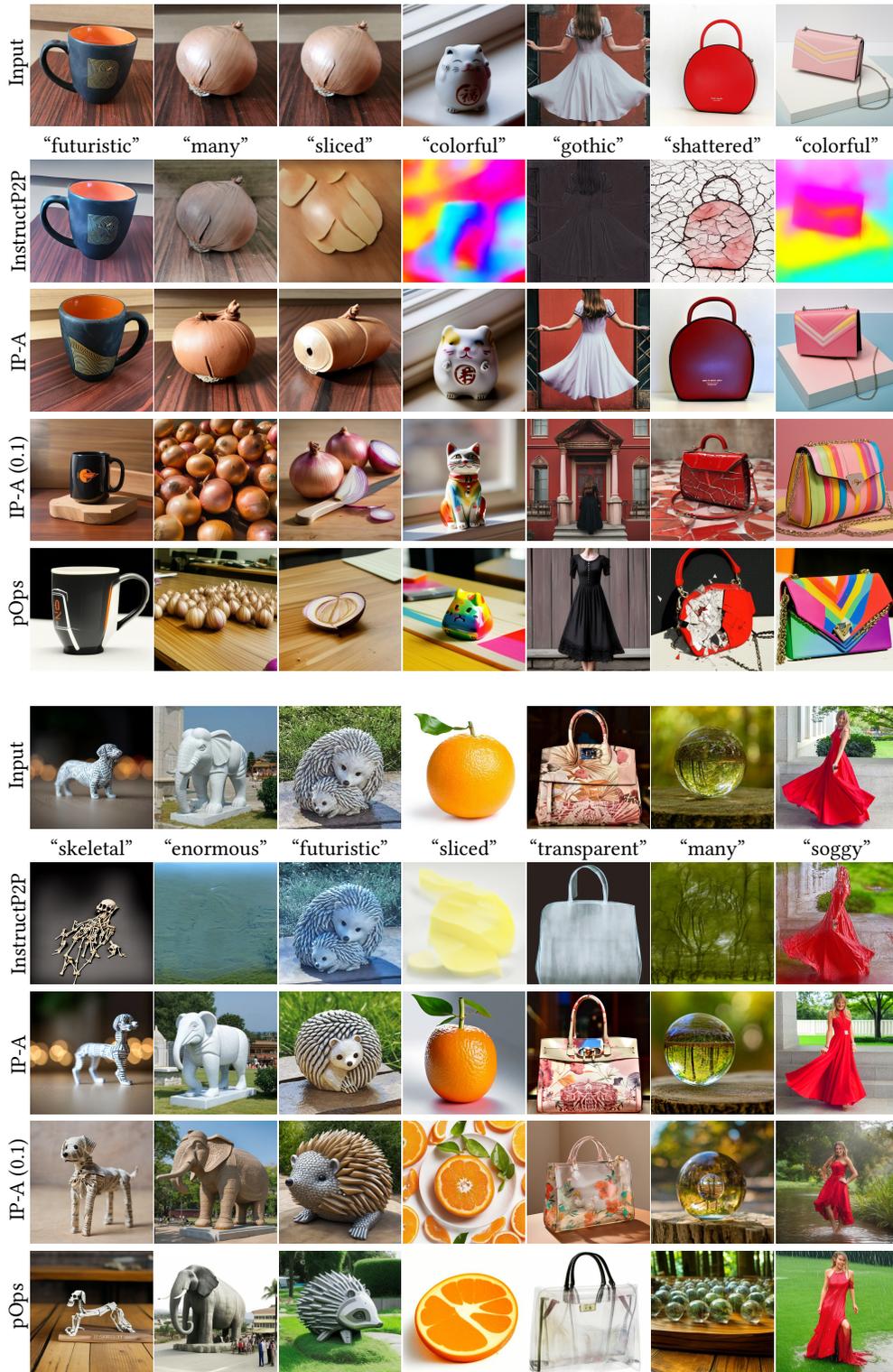


Fig. 22. Additional qualitative comparison for the *pOps* instruct operator to alternative instruction-based editing approaches: InstructPix2Pix [Brooks et al. 2023] and two variants of IP-Adapter [Ye et al. 2023].

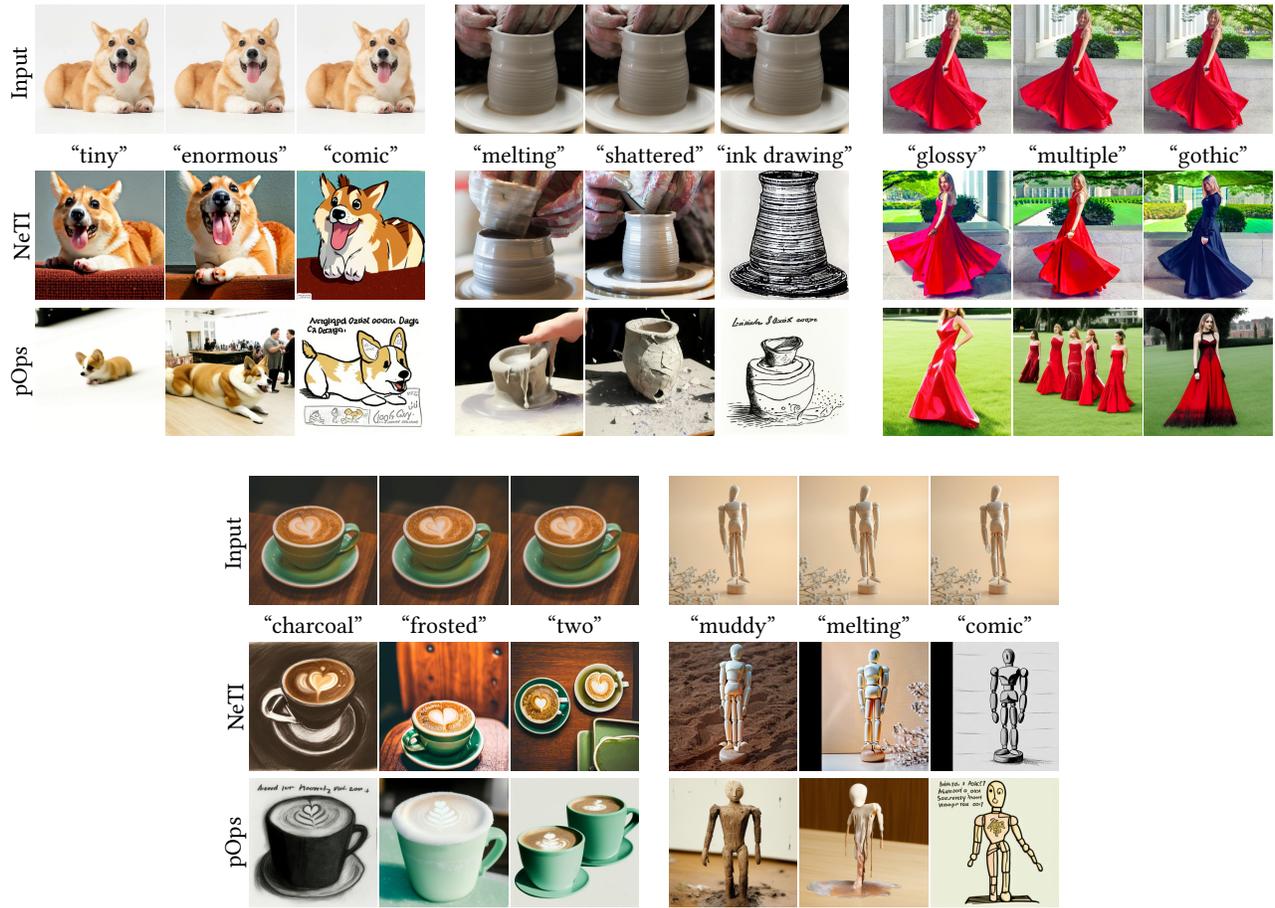
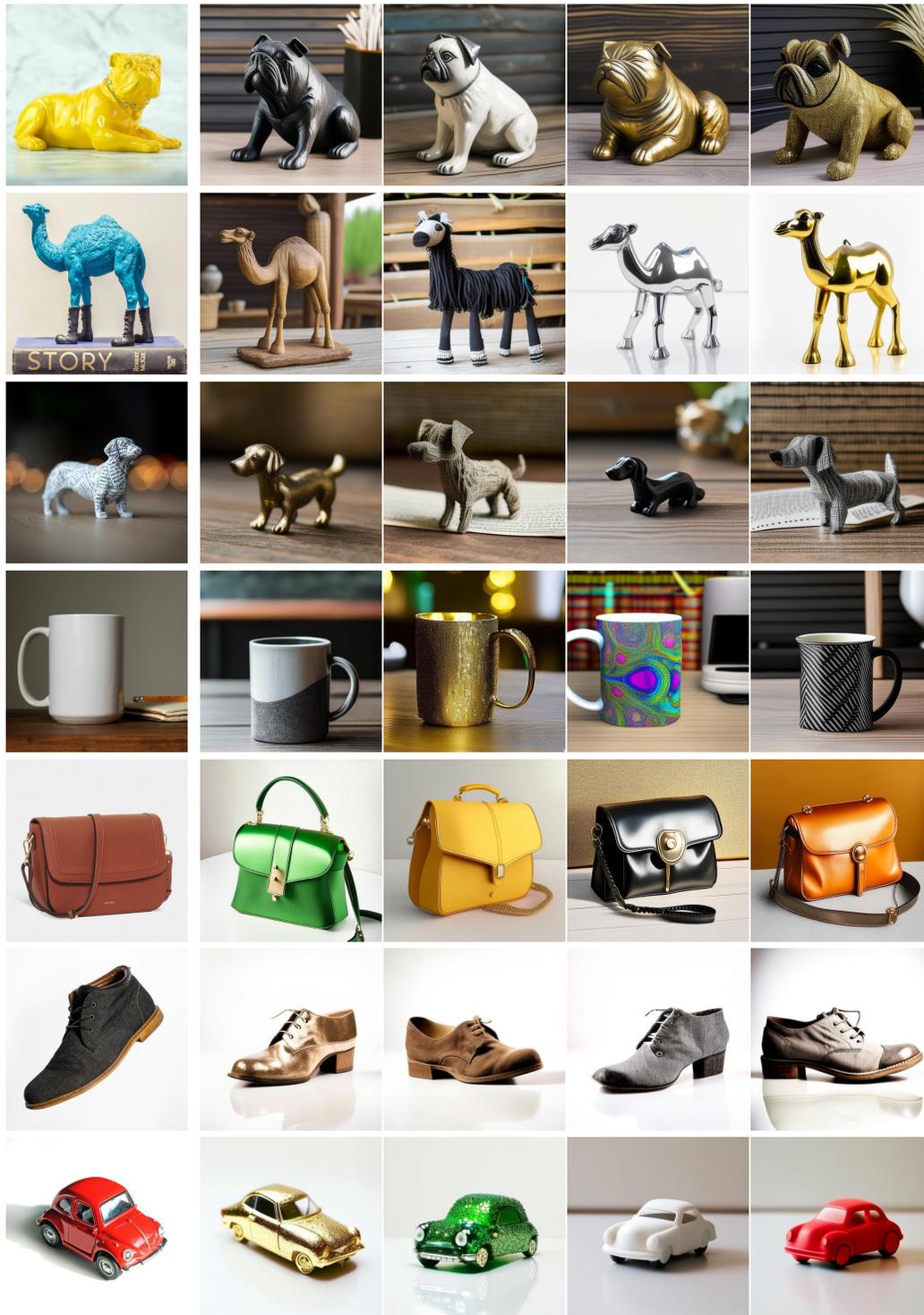


Fig. 23. Qualitative comparison for the pOps instruct operator compared to NeTI [Alaluf et al. 2023b], an optimization-based personalization technique.





Object

Generated textures using null input

Fig. 25. Results obtained by our texturing model when null inputs are passed in place of the texture input.

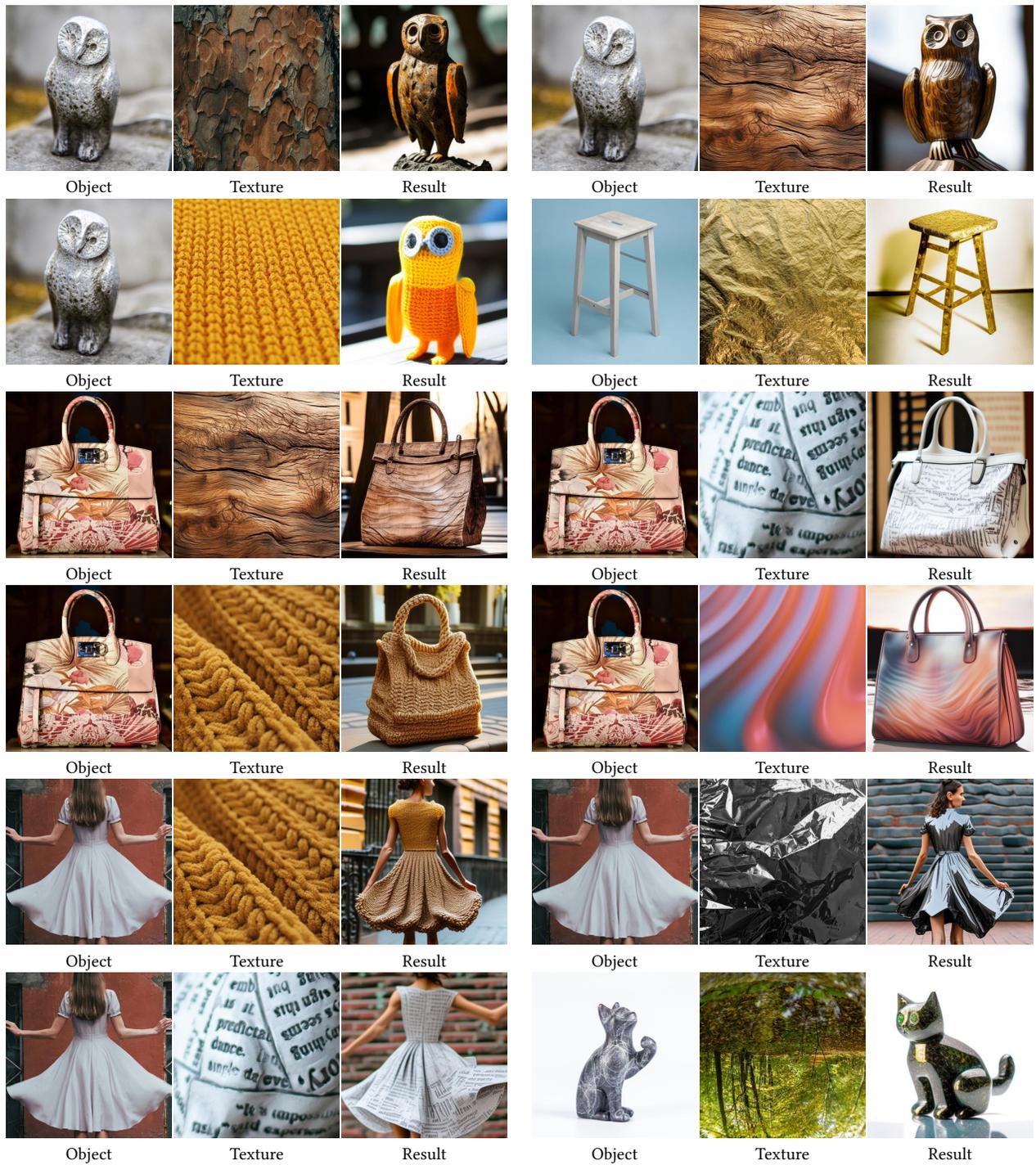


Fig. 26. Additional texturing results obtained by our *pOps* method.



Fig. 27. Additional texturing results obtained by our *pOps* method.

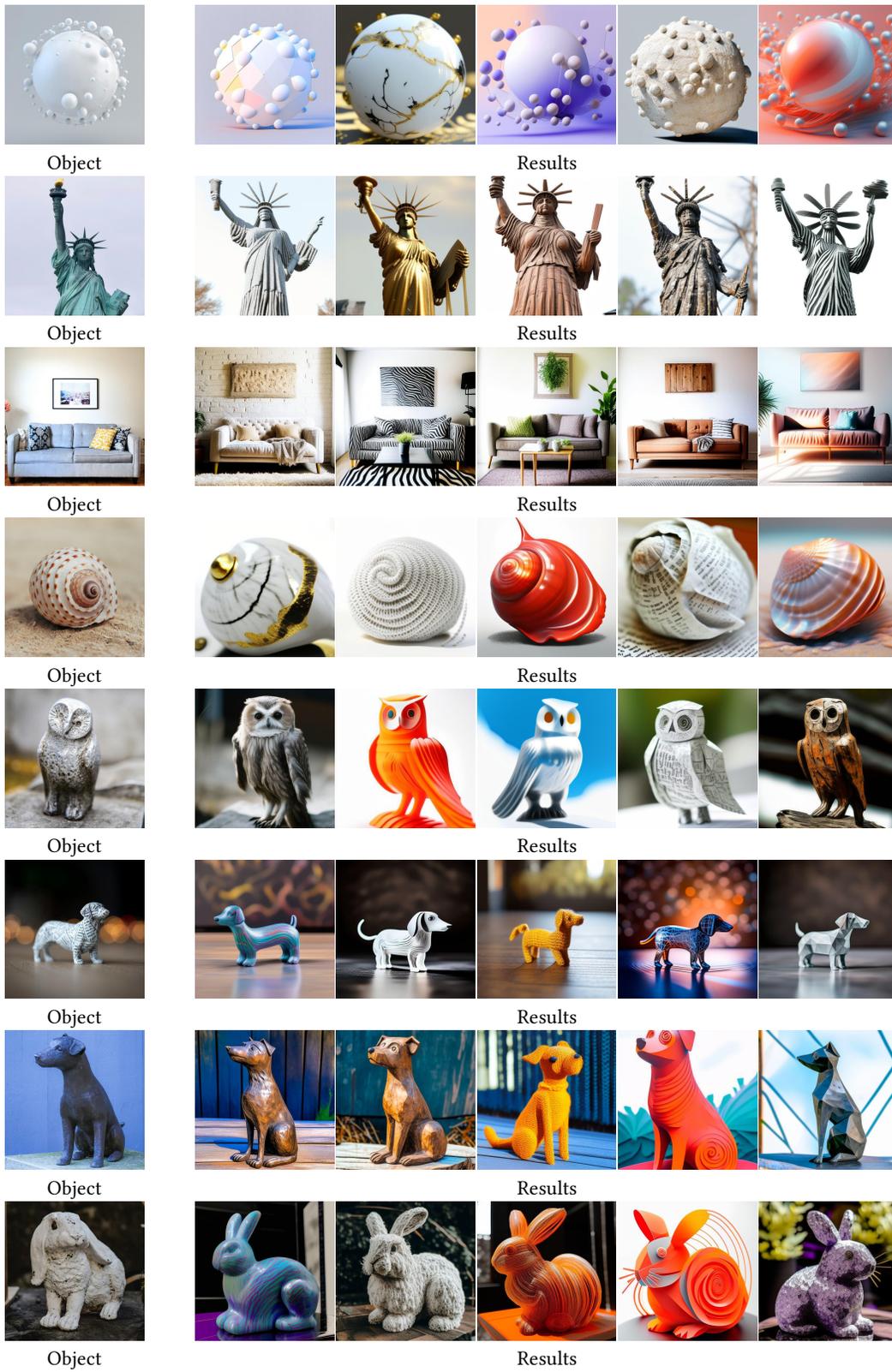


Fig. 28. Additional texturing operator results obtained by our *pOps* method.

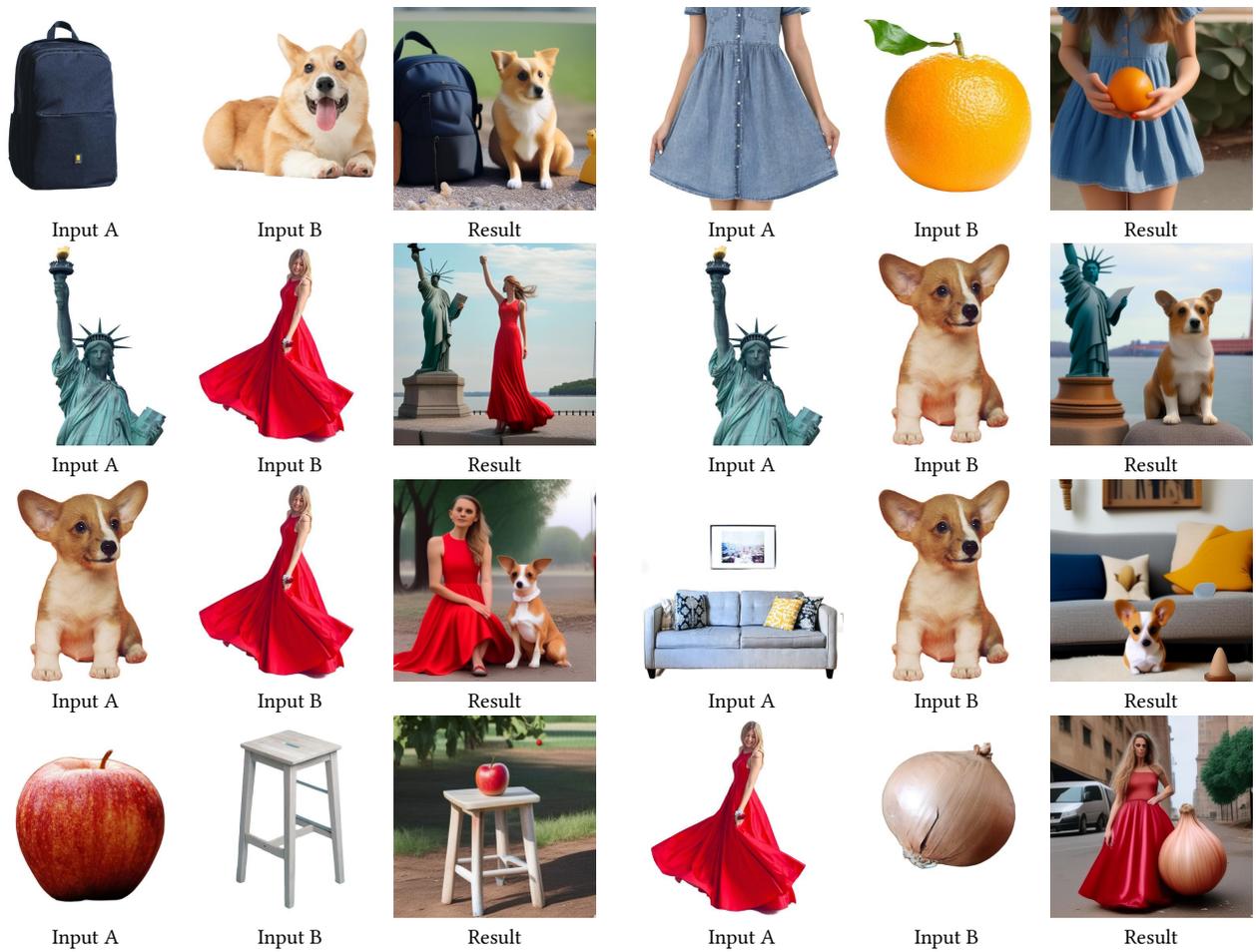


Fig. 29. Additional union results obtained by our *pOps* method.

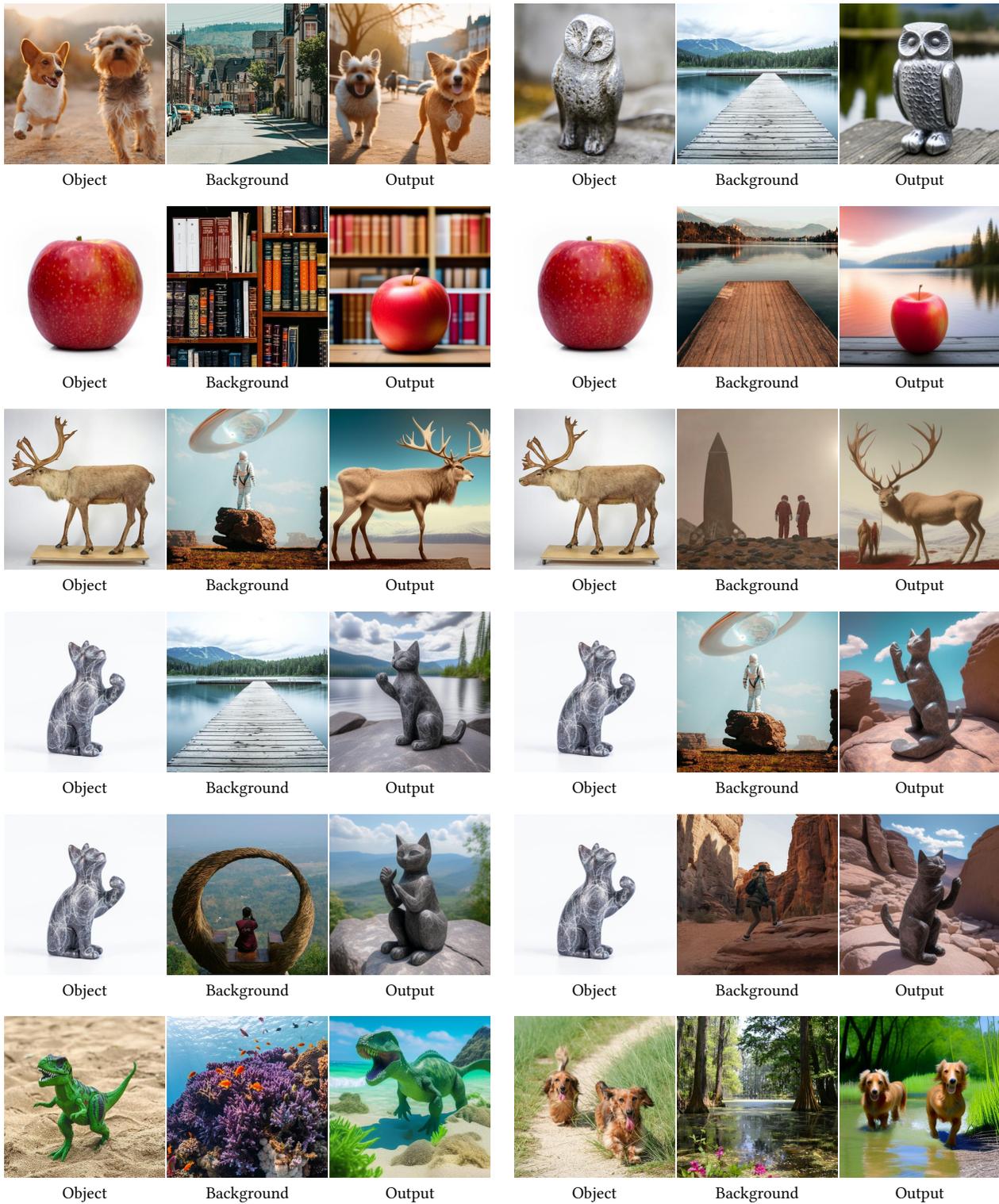


Fig. 30. Additional scene operator results obtained by our  $pOps$  method.

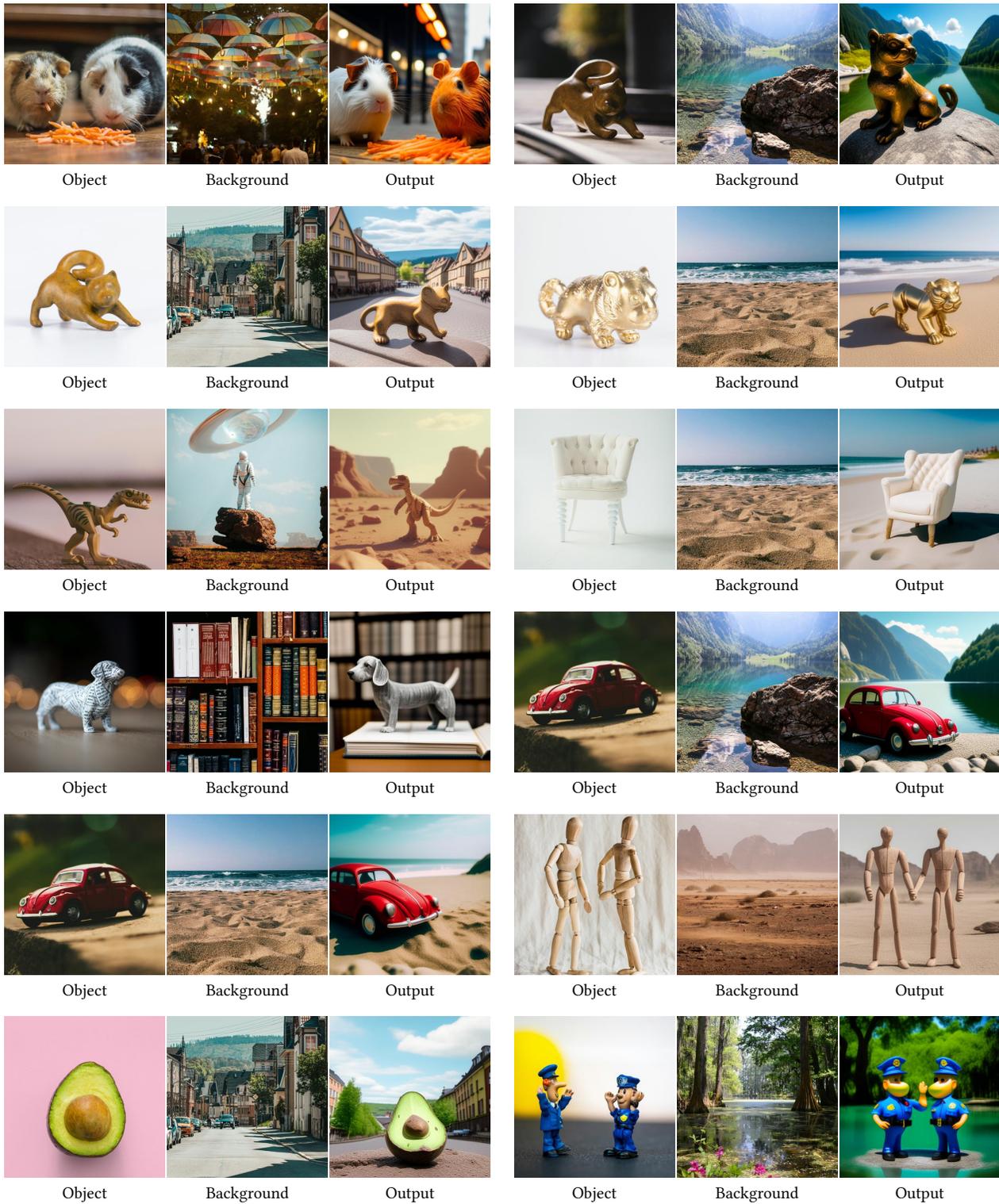


Fig. 31. Additional scene operator results obtained by our *pOps* method.

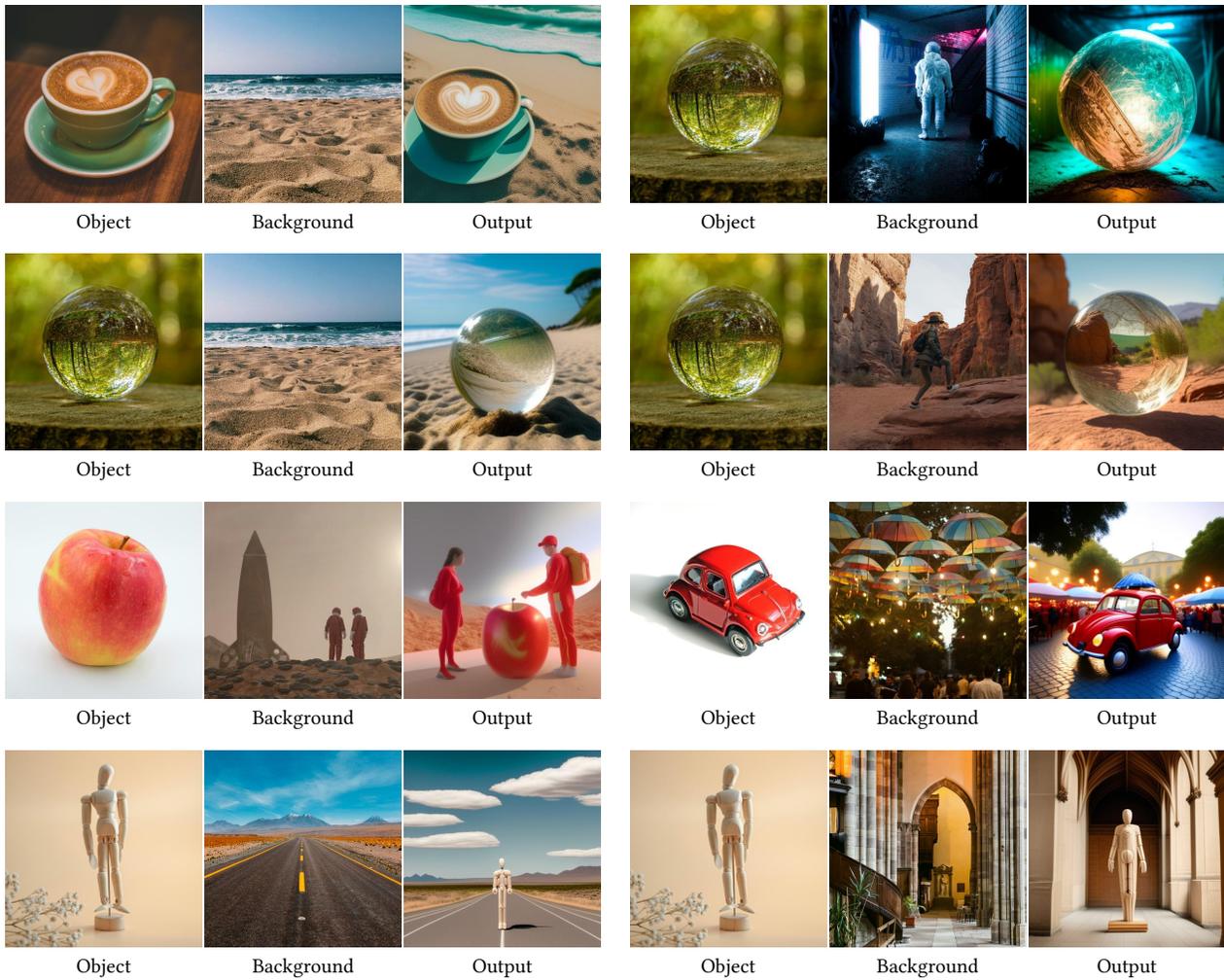


Fig. 32. Additional scene operator results obtained by our  $pOps$  method.





Fig. 34. Additional multi-image clothing composition results obtained by our *pOps* method.



Fig. 35. Additional multi-image composition results obtained by our *pOps* method.

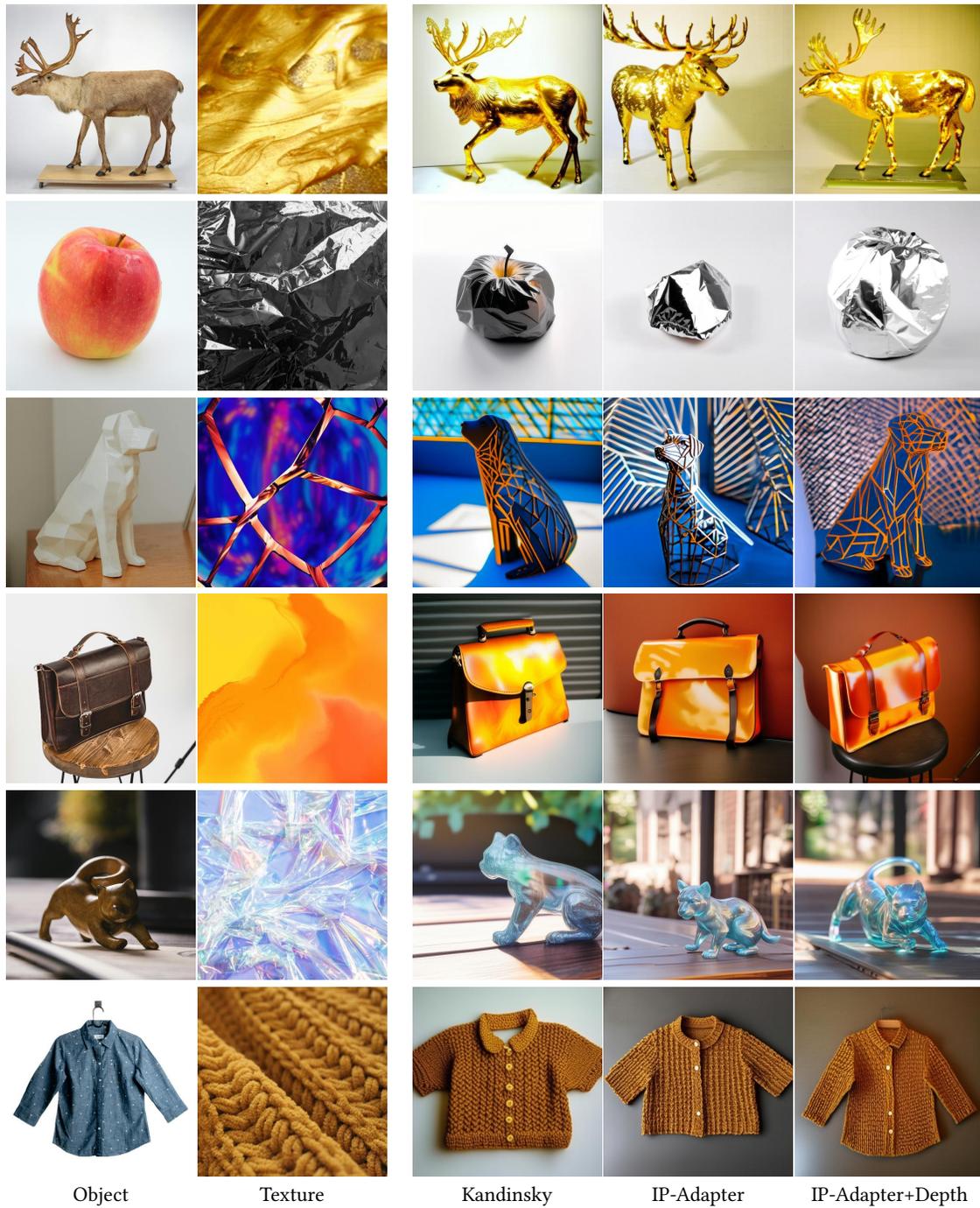


Fig. 36. **Different Renderers.** *pOps* outputs can be directly fed to either Kandinsky or IP-Adapter and incorporated alongside spatial conditions using ControlNet [Zhang and Agrawala 2023].

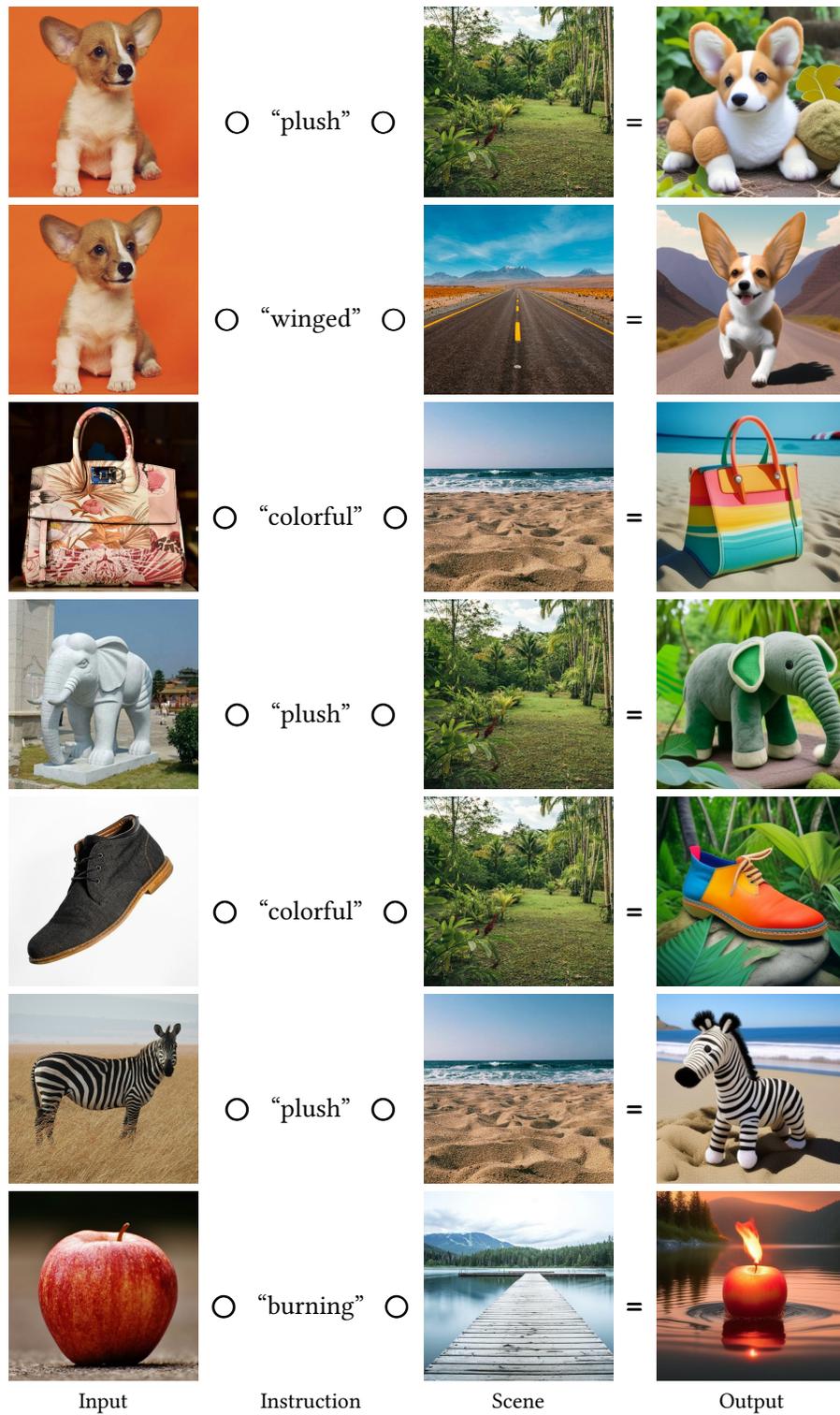


Fig. 37. Compositions of instruct and scene operators obtained by our pOps method.

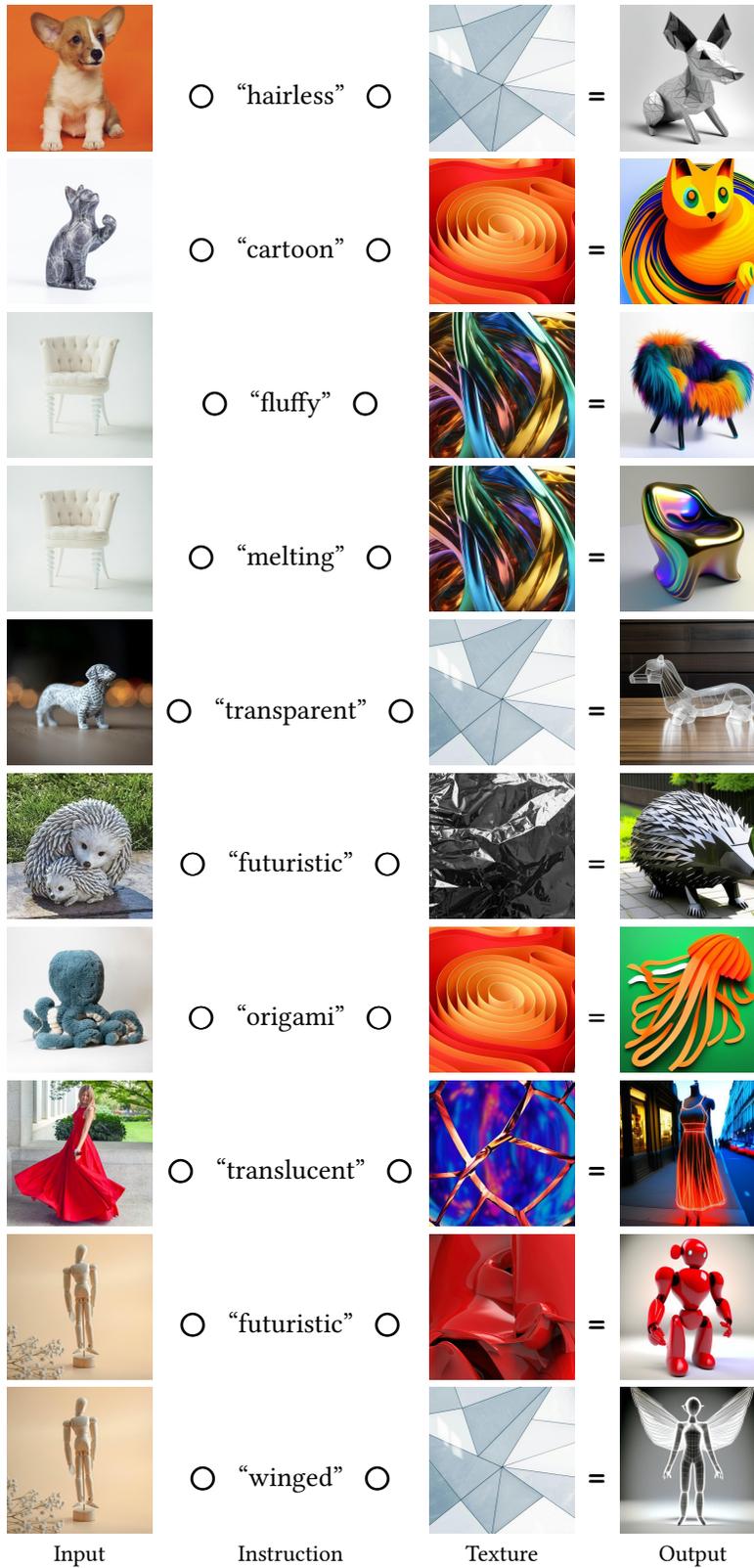


Fig. 38. Compositions of instruct and texturing operators obtained by our pOps method.

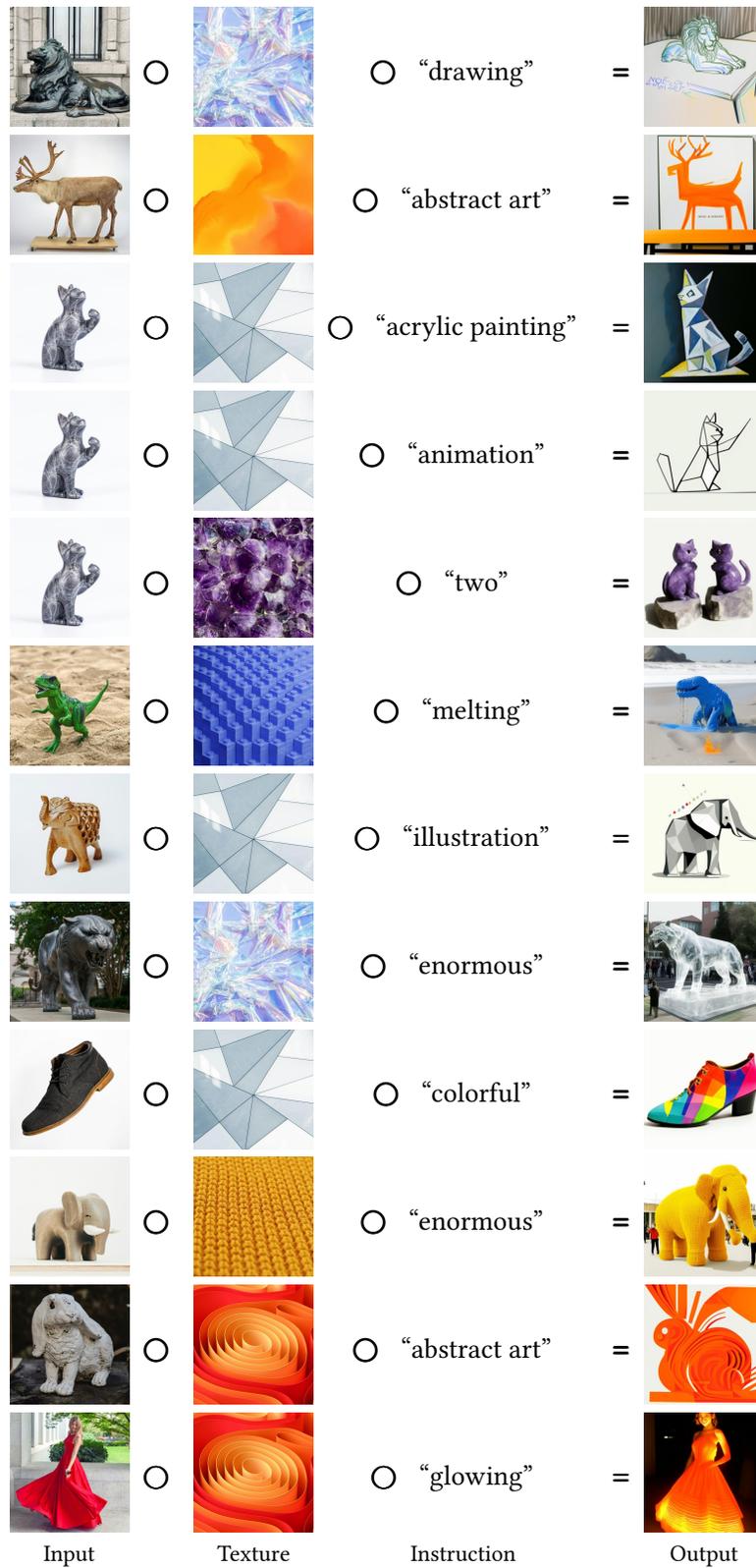


Fig. 39. Compositions of texturing and instruct operators obtained by our pOps method.