Break-A-Scene: Extracting Multiple Concepts from a Single Image

Omri Avrahami*
The Hebrew University of Jerusalem
Google Research
Jerusalem, Israel
omri.avrahami@mail.huji.ac.il

Kfir Aberman
Google Research
San Fransisco, USA
kfiraberman@gmail.com

Ohad Fried
Reichman University
Herzliya, Israel
ofried@runi.ac.il

Daniel Cohen-Or*
Tel Aviv University
Google Research
Tel Aviv, Israel
cohenor@gmail.com

Dani Lischinski*
The Hebrew University of Jerusalem
Google Research
Jerusalem, Israel
danix@mail.huji.ac.il

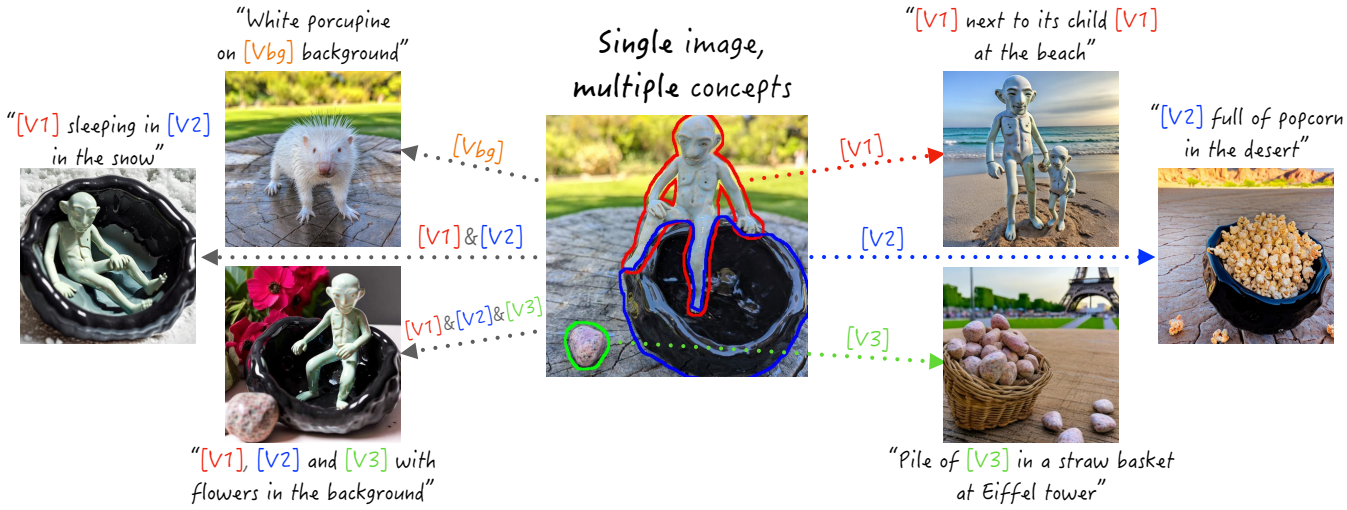

Figure 1: Break-A-Scene: Given a *single* image with *multiple* concepts, annotated by loose segmentation masks (middle), our method can learn a distinct token for each concept, and use natural language guidance to re-synthesize the individual concepts (right) or combinations of them (left) in various contexts.

ABSTRACT

Text-to-image model personalization aims to introduce a user-provided concept to the model, allowing its synthesis in diverse contexts. However, current methods primarily focus on the case of learning a *single* concept from *multiple* images with variations in backgrounds and poses, and struggle when adapted to a different scenario. In this work, we introduce the task of textual scene decomposition: given a *single* image of a scene that may contain *several* concepts, we aim to extract a distinct text token for each concept, enabling fine-grained control over the generated scenes. To this end, we propose augmenting the input image with masks that indicate the presence of target concepts. These masks can be provided by the user or generated automatically by a pre-trained segmentation model. We then present a novel two-phase customization process that optimizes a set of dedicated textual embeddings (handles), as well as the model weights, striking a delicate balance between accurately capturing the concepts and avoiding overfitting. We employ a masked diffusion loss to enable handles to generate their assigned concepts, complemented by a novel loss on cross-attention maps to prevent entanglement. We also introduce union-sampling, a training strategy aimed to improve the ability of combining multiple concepts in generated images. We use several automatic metrics to quantitatively compare our method against several baselines, and further affirm the results using a user study. Finally, we showcase several applications of our method.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Computer graphics**.

*Performed this work while working at Google

SA Conference Papers '23, December 12–15, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0315-7/23/12.
https://doi.org/10.1145/3610548.3618154

KEYWORDS

personalization, textual inversion, multiple concept extraction

ACM Reference Format:
Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. 2023. Break-A-Scene: Extracting Multiple Concepts from a Single Image. In *SIGGRAPH Asia 2023 Conference Papers (SA Conference Papers '23), December 12–15, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3610548.3618154

1 INTRODUCTION

Humans have a natural ability to decompose complex scenes into their constituent parts and envision them in diverse contexts. For instance, given a photo of a ceramic artwork depicting a creature seated on a bowl (Figure 1), one can effortlessly imagine the *same creature* in a variety of different poses and locations, or envision the *same bowl* in a new setting. However, today's generative models struggle when confronted with this type of task.

Recent works [Gal et al. 2022; Ruiz et al. 2023] suggested personalizing large-scale text-to-image models [Rombach et al. 2021; Saharia et al. 2022]: given *several* images of a *single* concept, they optimize newly-added dedicated text embeddings [Gal et al. 2022] or fine-tune the model weights [Ruiz et al. 2023] in order to enable synthesizing instances of this concept in novel contexts. These works initiated a vibrant research field, surveyed in more detail in Section 2 and summarized in Table 1.

In this work, we introduce the new scenario of *textual scene decomposition*: given a *single* image of a scene that may contain *multiple* concepts of different kinds, our goal is to extract a dedicated text token for each concept. This enables generation of novel images from textual prompts, featuring individual concepts or combinations of multiple concepts, as demonstrated in Figure 1.

The personalization task can be inherently ambiguous: it is not always clear which concepts we intend to extract/learn. Previous works [Gal et al. 2022; Ruiz et al. 2023] resolve this ambiguity by extracting a single concept at a time, utilizing several different images that depict the concept in different contexts. However, when switching to a single image scenario, other means are necessary to disambiguate the task. Specifically, we propose to augment the input image with a set of masks, indicating the concepts that we aim to extract. These masks may be loose masks provided by the user, or generated by an automatic segmentation method (e.g., [Kirillov et al. 2023]). However, as demonstrated in Figure 2, adapting the two main approaches, TI [Gal et al. 2022] and DB [Ruiz et al. 2023], to this setting reveals a reconstruction-editability tradeoff: while TI fails to accurately reconstruct the concepts in a new context, DB loses the ability to control the context due to overfitting.

In this work, we propose a novel customization pipeline that effectively balances the preservation of learned concept identity with the avoidance of overfitting. Our pipeline, depicted in Figure 3, consists of two phases. In the first phase, we designate a set of dedicated text tokens (handles), freeze the model weights, and optimize the handles to reconstruct the input image. In the second phase, we switch to fine-tuning the model weights, while continuing to optimize the handles.

We also recognize that in order to generate images exhibiting combinations of concepts, the customization process cannot be carried out separately for each concept. This observation leads us to introduce *union-sampling*, a training strategy that addresses this requirement and enhances the generation of concept combinations.

A crucial focus of our approach is on disentangled concept extraction, i.e., ensuring that each handle is associated with only a

Project page is available at: https://omriavrahami.com/break-a-scene/

Table 1: Scenarios of previous work on model personalization. Our method is the first to offer personalization of *multiple concepts* given a *single input image*. An extended version of this table, that also includes the concurrent works, is available in the supplementary materials.

Method	Single input image	Multi-concept output
Textual Inversion [Gal et al. 2022]	✗	✗
DreamBooth [Ruiz et al. 2023]	✗	✗
Custom Diffusion [Kumari et al. 2023]	✗	✓
ELITE [Wei et al. 2023]	✓	✗
E4T [Gal et al. 2023]	✓	✗
Ours	✓	✓

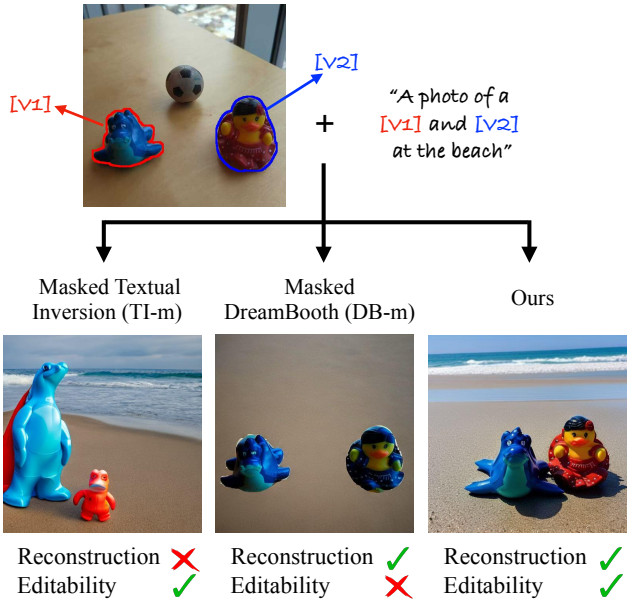

Figure 2: Reconstruction-editability tradeoff: Given a single input image along with masks, and the prompt "A photo of a $[v_1]$ and $[v_2]$ at the beach", (masked) Textual Inversion [Gal et al. 2022] generates an image of two objects on a beach, but fails to preserve their identities. (Masked) DreamBooth [Ruiz et al. 2023] preserves the identities well, but fails to place them on a beach. Our method is able to generate a convincing image of two objects on a beach, which closely resemble the objects in the input image.

single target concept. To achieve this, we employ a masked version of the standard diffusion loss, which guarantees that each custom handle can generate its designated concept; however, this loss does not penalize the model for associating a handle with multiple concepts. Our main insight is that we can penalize such entanglement by additionally imposing a loss on the cross-attention maps, known to correlate with the scene layout [Hertz et al. 2022]. This additional loss ensures that each handle attends only to the areas covered by its target concept.

We propose several automatic metrics for our task and use them to compare our method against the baselines. In addition, we conduct a user study and show that our method is also preferred by human evaluators. Finally, we present several applications of our method.

In summary, our contributions are: (1) we introduce the new task of textual scene decomposition, (2) propose a novel approach for this setting, which learns a set of disentangled concept handles, while balancing between concept fidelity and scene editability, and (3) propose several automatic evaluation metrics and use them, in addition to a user study, to demonstrate the effectiveness of our method.

2 RELATED WORK

Text-to-image synthesis. The field of text-to-image synthesis has seen immense progress in recent years. The initial approaches utilized RNNs [Mansimov et al. 2016], GANs [Reed et al. 2016; Xu et al. 2018; Zhang et al. 2017, 2018] and transformers [Gafni et al. 2022; Ramesh et al. 2021]. However, diffusion-based models [Ho et al. 2020; Sohl-Dickstein et al. 2015; Song et al. 2020; Song and Ermon 2019] emerged as superior for text-to-image generation [Chang et al. 2023; Ramesh et al. 2022; Rombach et al. 2021; Saharia et al. 2022; Yu et al. 2022].

Alongside these advancements, text-driven image editing has emerged, enabling global [Brooks et al. 2023; Crowson et al. 2022; Kwon and Ye 2022; Meng et al. 2021; Patashnik et al. 2021; Tumanyan et al. 2023; Valevski et al. 2022] and local manipulations [Avrahami et al. 2023a, 2022; Bar-Tal et al. 2022; Bau et al. 2021; Couairon et al. 2022; Kawar et al. 2023; Nichol et al. 2021; Patashnik et al. 2023; Sheynin et al. 2022; Wang et al. 2023]. In addition, diffusion models have also been employed for video generation [Ho et al. 2022; Singer et al. 2022], video editing [Molad et al. 2023], scene generation [Avrahami et al. 2023b; Bar-Tal et al. 2023], mesh texturing [Richardson et al. 2023], typography generation [Iluz et al. 2023], and solving inverse problems [Horwitz and Hoshen 2022; Saharia et al. 2021a,b].

Cross-attention. Prompt-to-prompt [Hertz et al. 2022] utilizes cross-attention maps in text-to-image diffusion models for manipulating generated images, later extended to handle real images through inversion [Mokady et al. 2023]. Attend-and-excite [Chefer et al. 2023] use cross-attention maps as an explainability-based technique [Chefer et al. 2020, 2021] to adjust text-to-image generations. In our work, we employ cross-attention maps to disentangle learned concepts; however, our work focuses on extracting textual handles from a scene and remixing them into completely novel scenes, rather than editing the input image.

Inversion. In the realm of generative models, *inversion* [Xia et al. 2021] is the task of finding a code within the latent space of a generator [Goodfellow et al. 2014; Karras et al. 2019, 2020] that faithfully reconstructs a given image. Inversion may be accomplished via direct optimization of the latent code [Abdal et al. 2019, 2020; Zhu et al. 2020a] or by training a dedicated encoder [Alaluf et al. 2021; Pidhorskyi et al. 2020; Richardson et al. 2020; Tov et al. 2021; Zhu et al. 2020b]. PTI [Roich et al. 2021] follows the latent optimization with refinement of the model weights [Bau et al. 2019]. In this study, we also employ a two-stage approach wherein we first optimize only the textual embeddings of the target concepts, followed by jointly training the embeddings and the model weights.

Personalization. The task of *personalization* aims to identify a user-provided concept that is not prevalent in the training data for discriminative [Cohen et al. 2022] or generative [Nitzan et al. 2022] tasks. Textual Inversion (TI) [Gal et al. 2022], and DreamBooth (DB) [Ruiz et al. 2023] are two seminal works that address personalization of text-to-image models: given *several* images of a *single* visual concept, they learn to generate this concept in different contexts. TI introduces a new learnable text token and optimizes it to reconstruct the concept using the standard diffusion loss, while keeping the model weights frozen. DB, on the other hand, reuses an existing rare token, and fine-tunes the model weights to reconstruct the concept. Custom Diffusion [Kumari et al. 2023] fine-tunes only a subset of the layers, while LoRA [Hu et al. 2021; Ryu 2022] restricts their updates to rank 1. Perfusion [Tewel et al. 2023] also performs a rank 1 update along with an attention key locking mechanism.

Concurrently with our work, SVDiff [Han et al. 2023] introduces an efficient personalization method in the parameter space based on singular-value decomposition of weight kernels. They also propose a mixing and unmixing regularization that enables generating two concepts next to each other. In contrast to our method, SVDiff requires *several* images for each of the concepts, while we operate on a *single* image containing *multiple* concepts. Furthermore, SVDiff's automatic augmentation allows for the placement of two objects side by side, while our method enables arbitrary placement of up to four objects.

Most recently, fast personalization methods were introduced that employ dedicated encoders [Chen et al. 2023; Gal et al. 2023; Jia et al. 2023; Shi et al. 2023; Wei et al. 2023] and can also handle a single image. Among these, only ELITE [Wei et al. 2023] is publicly available, and we include it in our comparisons in Section 4.1. XTI [Voynov et al. 2023] extends TI to utilize a richer inversion space. As shown in Table 1, our approach stands out from the existing personalization methods by addressing the challenge of coping with *multiple* concepts within a *single* image. To the best of our knowledge, this is the first work to tackle this task.

3 METHOD

Given a single input image I and a set of N masks $\{M_i\}_{i=1}^{N}$, indicating concepts of interest in the image, we aim to extract N *textual handles* $\{v_i\}_{i=1}^{N}$, s.t. the ith handle, v_i, represents the concept indicated by M_i. The resulting handles can then be used in text prompts to guide the synthesis of new instances of each concept, or novel combinations of several concepts, as demonstrated in Figure 1.

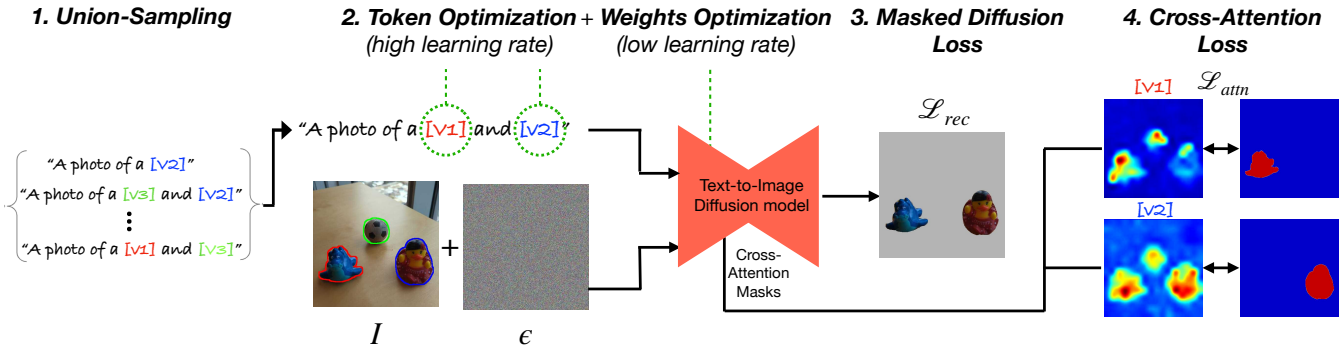

Figure 3: Method overview: our method consists of four key components: (1) in order to train the model to support different combinations of generated concepts, we employ a *union-sampling* mechanism, where a random subset of the tokens is sampled each time. In addition, (2) in order to avoid overfitting, we use a *two-phase training regime*, which starts by optimizing only the newly-added tokens, with a high learning rate, and in the second phase we also train the model weights, using a lower learning rate. A *masked diffusion loss* (3) is used to reconstruct the desired concepts. Finally, (4) in order to encourage disentanglement between the learned concepts, we use a novel *cross-attention loss.*

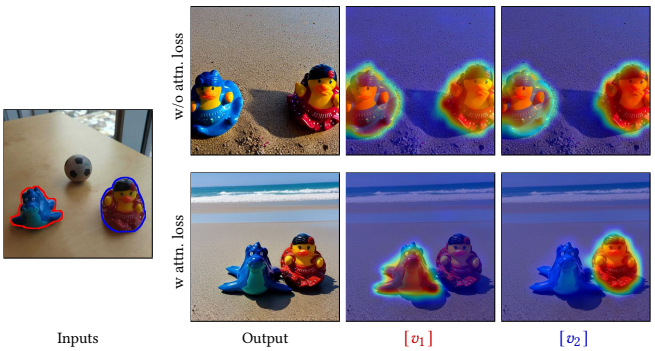

Figure 4: Cross-attention loss: given the input scene and the text prompt "a photo of $[v_1]$ and $[v_2]$ at the beach", we visualize the cross-attention maps of the generated images by our method with only the masked reconstruction loss of Equation (1) (top row), and after adding the cross-attention loss of Equation (2) (bottom row). Adding the cross-attention loss encourages each of the handles $[v_1]$ and $[v_2]$ to attend only to its corresponding concept, which results with a disentangled concepts' generation.

Attempting to adapt TI or DB to extraction of multiple concepts from a single image (by using masks, as explained in Section 4), reveals an inherent reconstruction-editability tradeoff. As demonstrated in Figure 2, TI enables embedding the extracted concepts in a new context, but fails to faithfully preserve their identity, while fine-tuning the model in DB captures the identity, at the cost of losing editability, to a point of failing to comply with the guiding text prompt. We observe that optimizing only individual tokens is not expressive enough for good reconstruction, while fine-tuning the model using a single image is extremely prone to overfitting. In this work, we strive for a "middle ground" solution that would combine the best of both worlds, i.e., would be able to capture the identity of the target concepts without relinquishing editability. Our approach combines four key components, as depicted in Figure 3 and described below.

Balancing between reconstruction and editability: We optimize both the text embeddings and the model's weights [Ryu 2022], but do so in two different phases. In the first phase, the model is frozen, while the text embeddings corresponding to the masked concepts are optimized [Gal et al. 2022] using a high learning rate. Thus, an initial approximate embedding is achieved quickly without detracting from the generality of the model, which then serves as a good starting point for the next phase. In the second phase, we unfreeze the model weights and optimize them along with the text tokens, using a significantly lower learning rate. This gentle fine-tuning of the weights and the tokens enables faithful reconstruction of the extracted concepts in novel contexts, with minimal editability degradation.

Union-sampling: We further observe that if the above process considers each concept separately, the resulting customized model struggles to generate images that exhibit a combination of several concepts (see Figure 7 and Section 4.1). Thus, we propose *union-sampling* for each of the two optimization phases. Specifically, we start by designating an initial textual embedding (handle) v_i for each concept indicated by mask M_i. Next, at each training step, we randomly select a *subset* of $k \leq N$ concepts, $s = \{i_1, \ldots, i_k\} \subseteq [N]$, and construct a text prompt "a photo of $[v_{i_1}]$ and $\ldots$ $[v_{i_k}]$". The optimization losses described below are then computed with respect to the union of the corresponding masks, $M_s = \bigcup M_{i_k}$.

Masked diffusion loss: The handles (and the model weights, in the second phase) are optimized using a masked version of the standard diffusion loss [Ryu 2022], i.e., by penalizing only over the pixels covered by the concept masks:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{z,s,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon \odot M_s - \epsilon_\theta(z_t, t, p_s) \odot M_s\|_2^2\right], \quad (1)$$

where z_t is the noisy latent at time step t, p_s is the text prompt, M_s is the union of the corresponding masks, ϵ is the added noise, and, ϵ_θ is the denoising network. Using the masked diffusion loss

in pixel space encourages the process to faithfully reconstruct the concepts. However, no penalty is imposed for associating a single handle with multiple concepts, as demonstrated in Figure 7. Thus, with this loss alone, the resulting handles fail to cleanly separate between the corresponding concepts.

In order to understand the source of this issue, it is helpful to examine the cross-attention maps between the learned handles and the generated images, as visualized in Figure 4 (top row). It may be seen that both handles $[v_1]$ and $[v_2]$ attend to the union of the areas containing the two concepts in the generated image, instead of each handle attending to just one concept, as we would have liked.

Cross-Attention loss: We therefore introduce another loss term that encourages the model to not only reconstruct the pixels of the learned concepts, but also ensures that each handle attends only to the image region occupied by the corresponding concept. Specifically, as illustrated in Figure 3 (right), we utilize the cross-attention maps for the newly-added tokens and penalize their MSE deviation from the input masks. Formally, we add the following term to loss in both training phases:

$$\mathcal{L}_{\text{attn}} = \mathbb{E}_{z,k,t}\left[\|CA_\theta(v_i, z_t) - M_{i_k}\|_2^2\right], \tag{2}$$

where $CA_\theta(v_i, z_t)$ is the cross-attention map between the token v_i and the noisy latent z_t. The cross attention maps are calculated over several layers of the denoising UNet model (for more details, please see the supplementary material). Thus, the total loss used is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{attn}}\mathcal{L}_{\text{attn}}, \tag{3}$$

where $\lambda_{\text{attn}} = 0.01$. As can be seen in Figure 4 (bottom row), the addition of $\mathcal{L}_{\text{attn}}$ to the loss succeeds in ensuring that $[v_1]$ and $[v_2]$ attend to two distinct regions, corresponding to the appropriate spatial locations in the generated image.

4 EXPERIMENTS

This section begins by adapting current text-to-image personalization methods to suit our single-image problem setting, followed by a qualitative comparison with our method. Next, we establish an automatic pipeline to evaluate the effectiveness of our method and compare it quantitatively to the baseline methods. Additionally, a user study is conducted to substantiate the claim that our method outperforms the baselines. Finally, we explore several applications, demonstrating the versatility and usefulness of our approach.

4.1 Comparisons

Existing personalization methods, such as DreamBooth [Ruiz et al. 2023] and Textual Inversion [Gal et al. 2022], take multiple images as input, rather than a single image with masks indicating the target concepts. Applying such methods to a single image without such indication results in tokens that do not necessarily correspond to the concepts that we wish to learn. For more details and examples, please see the supplementary material.

Thus, in order to conduct a meaningful comparison with these previous methods, we first adapt them to our problem setting. This is achieved by converting a single input image with several concept masks into a small collection of image-text pairs, as shown in Figure 6. Specifically, each pair is constructed by randomly choosing a subset of concepts $i_1, \ldots, i_k$, and placing them on a random solid color background with a random flip augmentation. The text prompt accompanying each such image is "a photo of $[v_{i_1}]$ and ... $[v_{i_k}]$". We refer to DB and TI trained on such image collections as DB-m and TI-m, respectively.

Another personalization approach, Custom Diffusion [Kumari et al. 2023] (CD), optimizes only the cross-attention weights of the denoising model, as well as a newly-added text token. We adapt CD to our problem setting using the same approach as above, and refer to the adapted version as CD-m. In addition, ELITE [Wei et al. 2023], trains encoders on a *single image* to allow fast personalization, and also supports input masks. We use the official implementation of ELITE to compare it with our method.

Qualitative comparisons. We start with a qualitative comparison between our method and the baselines. As demonstrated in Figure 9, TI-m and CD-m are able to generate images that follow the text prompt, but struggle with preserving the concept identities. DB-m preserves the identities well, but is not able to generate an image that complies with the rest of the prompt. ELITE preserves the identities better than TI-m and CD-m, but the reconstruction is still not faithful to the input image, especially when trying to generate more than one concept. Finally, our method is able to generate images that preserve the identity as well as follow the text prompt, and we demonstrate this ability with up to four different concepts in a single image.

Quantitative comparisons. In order to evaluate our method and the baselines quantitatively, we propose an automatic pipeline to generate a large number of inputs. As a source for these inputs, we use the COCO dataset [Lin et al. 2014], which contains images along with their instance segmentation masks. We crop COCO images into a square shape, and filter only those that contain at least two segments of distinct "things" type, with each segment occupying at least 15% of the image. We also filter out concepts from COCO classes that do not have individual identities (e.g., oranges). Then, in order to create a larger dataset, we pair each of these inputs with a random text prompt and a random number of tokens, yielding a total number of 5400 image-text pairs per baseline. For more details and examples, please read the supplementary material.

For each of the baselines TI-m, CD-m, and DB-m, we convert each input image and masks into a small image collection, as described earlier. For ELITE, we used the official implementation that supports an input mask. Next, we generate the results for each of the input image-text pairs with all the baselines, as well as with our method.

We employ two evaluation metrics: prompt similarity and identity similarity. Prompt similarity measures the degree of correspondence between the input text prompt and the generated image. Specifically, we utilize the standard CLIP similarity metric [Radford et al. 2021], i.e., the cosine between the normalized CLIP embeddings of the input prompt and the generated image. In each input prompt, the special $[v_i]$ tokens have been replaced with the text describing the corresponding class (e.g., "a photo of a cat at the beach" instead of "a photo of $[v_1]$ at the beach", which was used to create the image).

For the identity similarity metric, we must adapt the standard approach in order to deal with multiple subjects. A direct comparison

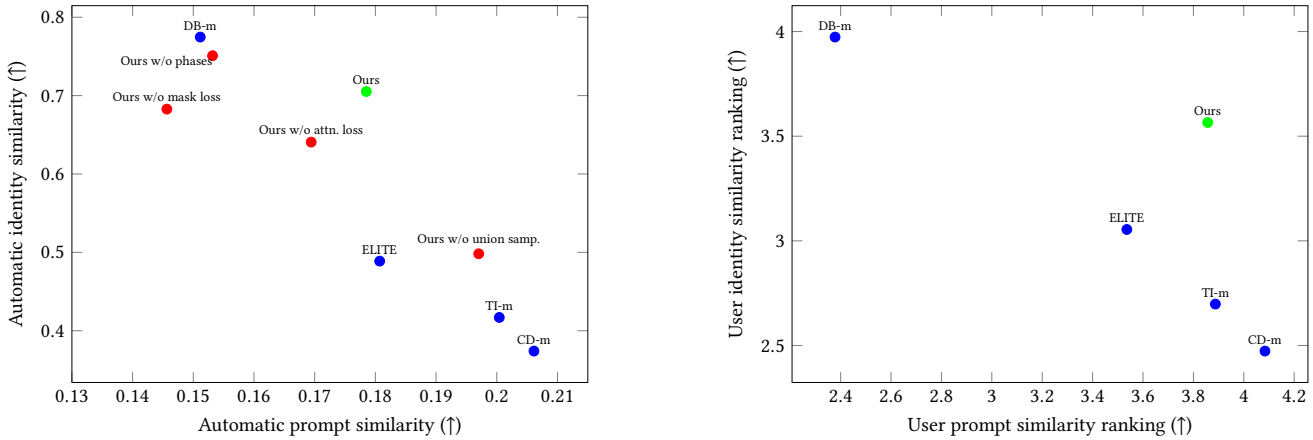

Figure 5: Quantitative comparison: (Left) A scatter plot of different personalization methods in terms of prompt similarity and identity similarity, generated as described in Section 4. DB-m preserves the identities, while compromising the prompt similarity. TI-m and CD-m follow the prompt, while sacrificing identity similarity. Our method lies on the Pareto front by balancing between the two extremes. We also plot ablated versions of our method: removing the first phase of our method reduces prompt similarity, removing the masked diffusion loss significant degrades prompt similarity, while removing the cross-attention loss or union sampling both degrade identity similarity. (Right) A scatter plot of human rankings of identity and prompt similarities (collected using Amazon Mechanical Turk) exhibits similar trends.

Figure 6: Baseline adaptation: Given a single image with concept masks, we construct a small collection of image-text pairs by selecting a random subset of tokens each time, creating a text prompt in the format of "A photo of $[v_x]$ and $[v_y]$ ...", masking out the background using the provided masks and applying a random solid background.

between the input image and the generated image is bound to be imprecise, because either image may contain multiple concepts: the input image contains all the concepts, while the generated one will typically contain a subset of them. Therefore, for each generated image, we compare a masked version of the input image (using the input mask from the COCO dataset) with a masked version of the generated image. We obtained the masked version of the generated image by leveraging MaskFormer [Cheng et al. 2021], a pre-trained image segmentation model.

In addition, following Ruiz et al. [2023], we chose to extract the image embeddings from the DINO model [Caron et al. 2021], as it was shown [Ruiz et al. 2023] to better capture the identity of objects, which aligns with the goals of personalization.

As demonstrated in Figure 5(left), there is an inherent trade-off between identity similarity and prompt similarity, with DB-m on one end, preserving the identity well, while sacrificing prompt similarity. TI-m and CD-m are on the other end of the spectrum, exhibiting high prompt similarity but low identity preservation. It may be seen that ELITE also struggles with preservation of identities. Our method lies on the Pareto front, balancing between the two requirements.

Ablation study. In addition, we conducted an ablation study, which includes removing the first phase (TI) of our method, removing the masked diffusion loss in Equation (1), removing the cross-attention loss in Equation (2), and training the model to reconstruct a single concept at each sample, instead of union-sampling. As seen in Figure 5(left), removing the first phase causes a significant degradation in prompt similarity, as the model tends to overfit. Removing the masked loss also causes a significant prompt similarity degradation, as the model tends to learn also the background of the original image, which may override the guiding text prompt. Removing the cross-attention loss yields a degradation of identity similarities, because the model does not learn to disentangle the concepts, as explained in Section 3. Finally, removing union-sampling degrades the ability of the model to generate images with multiple concepts, thereby significantly reducing the identity preservation score.

Figure 7 provides a visual comparison of the ablated cases. As can be seen, removing the first training phase causes the model to tend to ignore the target prompt. Removing the masked loss causes the model to extract elements from the background together with the masked concepts (note the vertical wooden poles present in the generated images). Removing the cross-attention loss causes the model to entangle between the concepts (the orange juice glass appears, even when the prompt only asks for the bear). When training without union sampling, the model struggles when asked generating more than one concept. For additional visual examples, please refer to the supplementary materials.

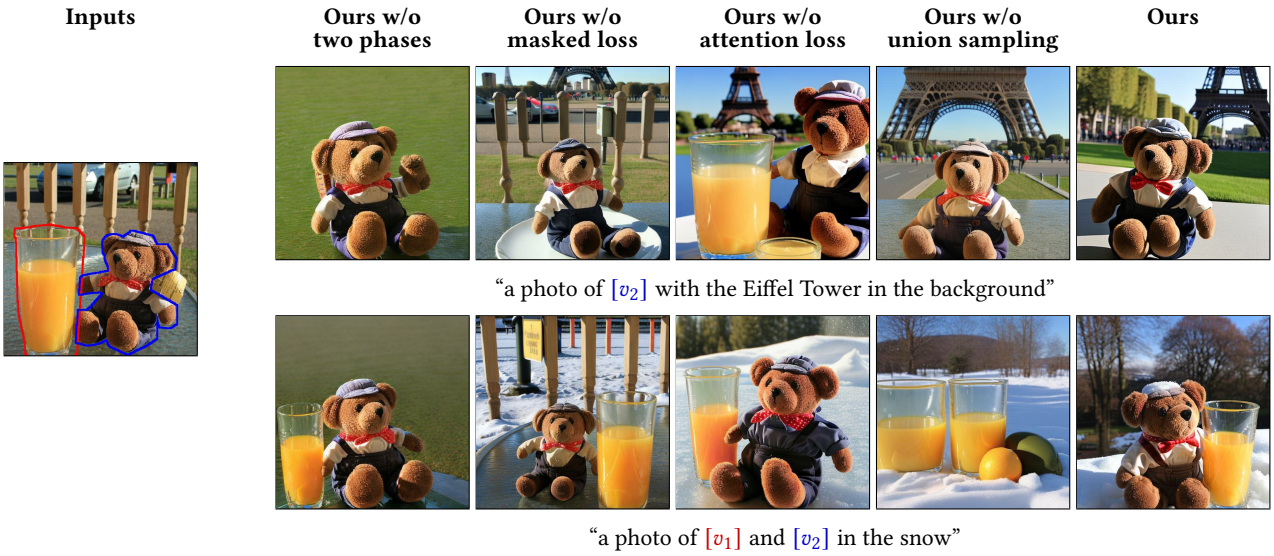

Figure 7: Qualitative ablation: we ablate our approach by removing the first training phase, removing the masked diffusion loss, removing the cross-attention loss, and sampling a single concept at a time. As can be seen, when removing the first training phase, the model overfits and fails to follow the guiding prompt, when removing the masked loss, the model tends to learn also the background. Without the cross-attention loss, the model tends to entangle the concepts or replicate one of them. Finally, when sampling a single concept at a time, the model struggles with generating images with multiple concepts.

User study. Lastly, we conducted a user study using the Amazon Mechanical Turk (AMT) platform. We chose a random subset of the automatically generated inputs from COCO, and asked the evaluators to rate the identity preservation and the prompt similarity of each result on a Likert scale of 1–5. When rating the prompt similarity, evaluators were presented with the input text prompt where the special $[v_i]$ tokens have been replaced with the text describing the corresponding class (as we did for the automatic prompt similarity metric). The results of our method and all the baselines were presented on the same page, and the evaluators were asked to rate each of the images. For identity preservation, we showed a masked version of the input image, containing only the object being generated, next to each of the results, and asked the evaluator to rank on the scale of 1–5 whether the images contain the same object as in the masked input image. For more details and statistical significance analysis, read the supplementary materials. As can be seen in Figure 5(right), the human rankings provide an additional evidence that our method lies on the Pareto front, balancing identity preservation and prompt similarity.

4.2 Applications

In Figure 10 we present several applications and use cases demonstrating the versatility of our method.

Image variations. Given a single image containing multiple concepts of interest, once these are extracted using our method, they can be used to generate multiple variations of the image by simply prompting the model with "a photo of $[v_1]$ and $[v_2]$ and ...". As demonstrated in Figure 10(a), the arrangement of the objects in the scene, as well as the background, are different in each generation.

Entangled concept separation. Given a single image with composite objects, one can decompose such objects into distinct concepts. For example, as shown in Figure 10(b), given a single image of a dog wearing a shirt, our method is able to separately learn the dog and the shirt concepts. Thus, it is possible to generate images of the dog without the shirt, or of a cat wearing that specific shirt. Note how the dog's body is not visible in the input image, yet the strong priors of the diffusion model enable it to generate a plausible body to go with the dog's head.

Background extraction. In addition to learning various foreground objects in the scene, we also learn the background as one of the visual concepts. The background mask is automatically defined as the complement of the union of all the input masks. As demonstrated in Figure 10(c), the user can extract the specific beach from the input image, and generate new objects on it. Please notice the correct water reflections of the newly generated objects. We used the same technique in Figure 1 (the white porcupine example). Note that this application is different from inpainting, as the model learns to generate variants of the background, e.g., the clouds in Figure 10(c) and the subtle background change in Figure 1.

Local image editing. Once concepts have been extracted, one may utilize an off-the-shelf text-driven local image editing method in order to edit other images, e.g., Blended Latent Diffusion [Avrahami et al. 2023a, 2022]. This is demonstrated in Figure 10(d): after extracting the concepts from the input scene of Figure 1, one may provide an image to edit, indicate the regions to be edited, and provide a guiding text prompt for each region. Then, by using Blended Latent Diffusion, we can embed the extracted concepts inside the indicated regions, while preserving the rest of the image. For more details on this approach, please refer to the supplementary material.

Figure 8: Limitations: our method suffers from several limitations: (a) in some cases, the model does not learn to disentangle between the lighting of the scene in the original single image and the learned concepts, s.t. the lighting become inconsistent with the target prompt. (b) In other cases, the model learns to entangle between the pose of the objects in the single input image and their identities, s.t. it is not able to generate them with different poses, even when explicitly being told to do so. (c) We found our method to work best when used to extract up to four concepts; when trying to extract more than that, our method tends to fail in learning the objects' identities. Credits: RebaSpike @ pixabay

This application is reminiscent of exemplar-based image editing methods [Song et al. 2022; Yang et al. 2023] with two key differences: (1) our single example image may contain multiple concepts, and (2) we offer an additional fine-grained textual control over each of the edited regions.

5 LIMITATIONS AND CONCLUSIONS

We found our method to suffer from the following limitations: (a) inconsistent lighting — because the input to our method is a single image, our method sometimes struggles with disentangling the lighting from the learned identities, e.g., the input image in Figure 8(a) was taken in broad daylight, and the model learns to generate the extracted concepts with daylight lighting, even when the user prompts it specifically with different environments (coral reef, dark cave and dark night). (b) Pose fixation — another problem that stems from the single input is that sometimes the model learns to entangle between the object pose and its identity, e.g., the input image in Figure 8(b) contains a dog looking upward with an open mouth, and the model generates the dog in this position in all the images, even when instructed specifically to refrain from doing so. (c) Underfitting of multiple concepts — we found that our method works best when given up to four concepts, e.g., the input in Figure 8(c) contains six objects, and the model struggles when learning that many identities. (d) Significant computational cost and parameter usage — our method takes about 4.5 minutes to extract the concepts from a single scene and to fine-tune the entire model. Incorporating recent faster approaches that are more parameter-efficient (e.g., Custom Diffusion) did not work, which limits the applicability of this approach in time-sensitive scenarios. Improving the model cost is an appealing direction for further research.

In conclusion, in this paper we address the new scenario of extracting multiple concepts from a single image. We hope that it will serve as a building block for the future of the field, as generative AI continues to evolve and push the boundaries of what is possible in the realm of creative expression.

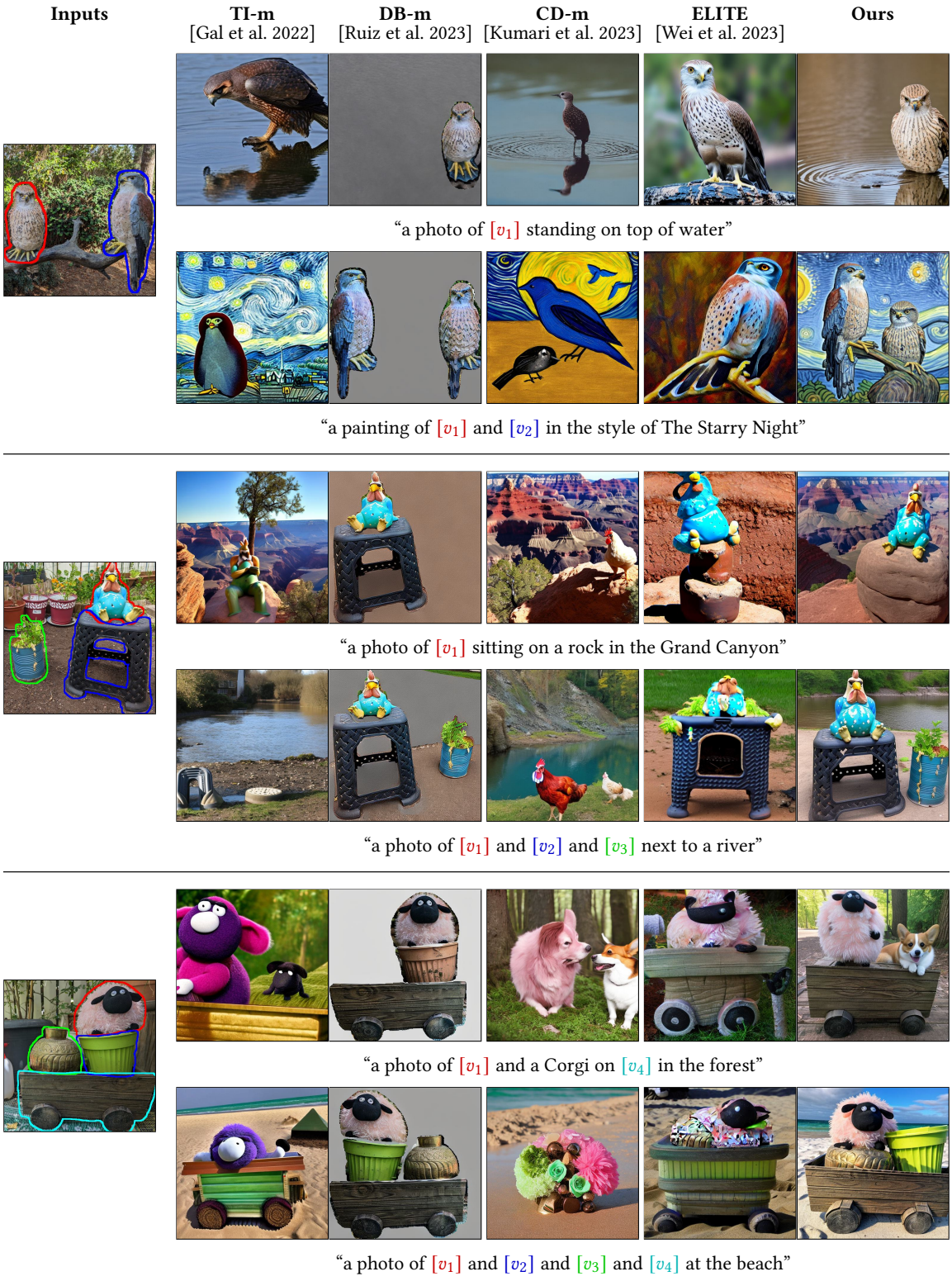

Figure 9: A qualitative comparison between several baselines and our method. TI-m and CD-m struggle with preserving the concept identities, while the images generated by DB-m effectively ignore the text prompt. ELITE preserves the identities better than TI-m/CD-m, but the concepts are still not recognizable enough, especially when more than one concept is generated. Finally, our method is able to preserve the identities as well as follow the text prompt, even when learning four different concepts (bottom row).

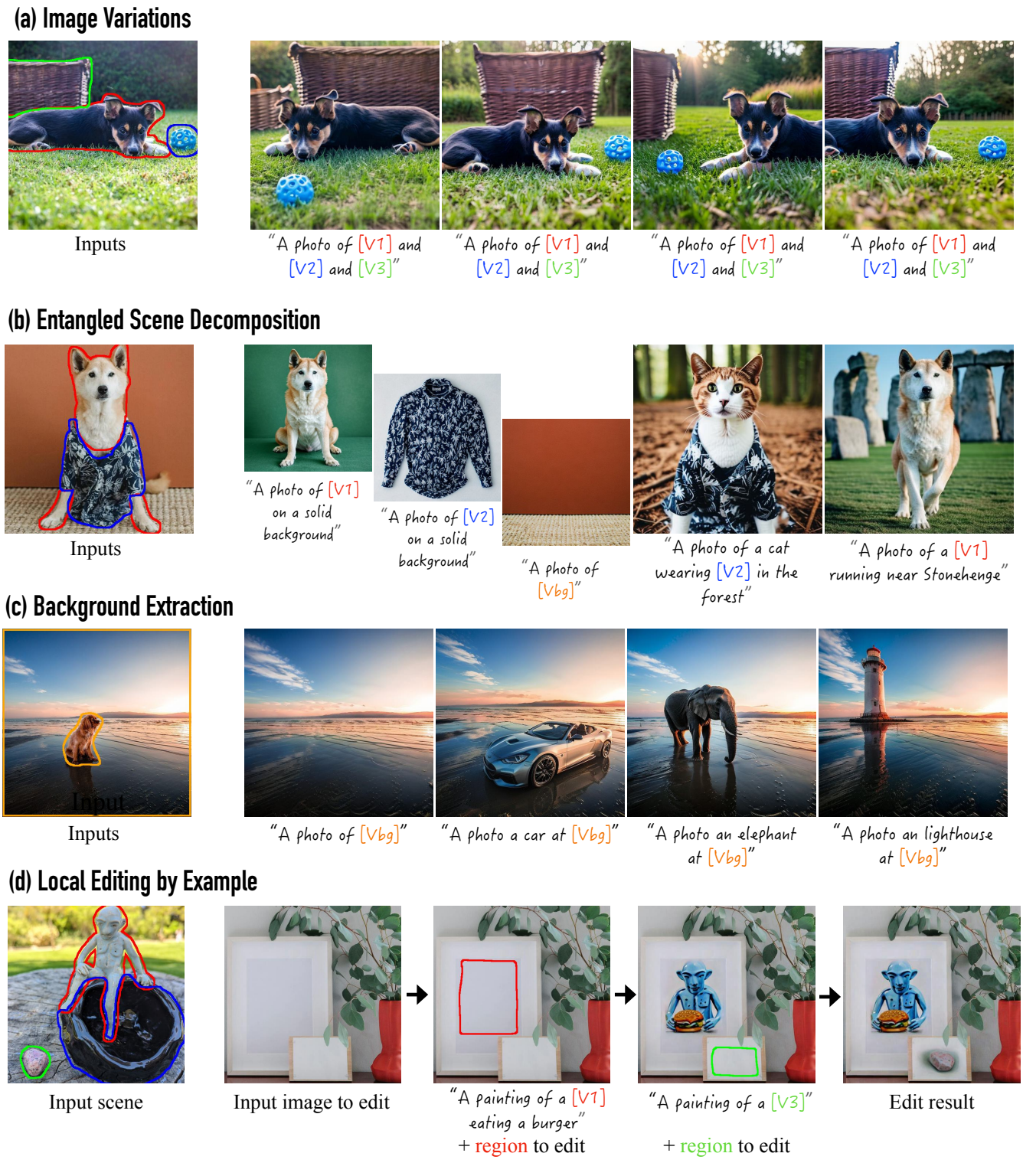

Figure 10: Applications: our method can be used for other downstream tasks, such as generating image variations, decomposing entangled concepts into their components, extracting the background from an existing scene, and locally editing an existing image using off-the-shelf tools [Avrahami et al. 2023a, 2022]. Credits: Magda Ehlers @ pexels / Sam Lion @ pexels / pixabay / Angela Roma @ pexels

REFERENCES

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2stylegan: How to embed images into the stylegan latent space?. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4432–4441.

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020. Image2stylegan++: How to edit the embedded images?. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8296–8305.

Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Haim Bermano. 2021. HyperStyle: StyleGAN Inversion with HyperNetworks for Real Image Editing. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 18490–18500. https://api.semanticscholar.org/CorpusID:244729249

Omri Avrahami, Ohad Fried, and Dani Lischinski. 2023a. Blended Latent Diffusion. *ACM Trans. Graph.* 42, 4, Article 149 (jul 2023), 11 pages. https://doi.org/10.1145/3592450

Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. 2023b. SpaText: Spatio-Textual Representation for Controllable Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18370–18380.

Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18208–18218.

Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. 2022. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*. Springer, 707–723.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. 2023. Multidiffusion: Fusing diffusion paths for controlled image generation. (2023).

David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. 2021. Paint by Word. arXiv:2103.10951 [cs.CV]

David Bau, Hendrik Strobelt, William S. Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. 2019. Semantic photo manipulation with a generative image prior. *ACM Transactions on Graphics (TOG)* 38 (2019), 1 – 11.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *CVPR*.

Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 9630–9640.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, José Lezama, Lu Jiang, Ming Yang, Kevin P. Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. 2023. Muse: Text-To-Image Generation via Masked Generative Transformers. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:255372955

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. *ACM Transactions on Graphics (TOG)* 42 (2023), 1 – 10. https://api.semanticscholar.org/CorpusID:256416326

Hila Chefer, Shir Gur, and Lior Wolf. 2020. Transformer Interpretability Beyond Attention Visualization. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 782–791.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 387–396.

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W. Cohen. 2023. Subject-driven Text-to-Image Generation via Apprenticeship Learning. *ArXiv* abs/2304.00186 (2023).

Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. 2021. Per-Pixel Classification is Not All You Need for Semantic Segmentation. In *Neural Information Processing Systems*.

Niv Cohen, Rinon Gal, Eli A Meirom, Gal Chechik, and Yuval Atzmon. 2022. "This is my unicorn, Fluffy": Personalizing frozen vision-language representations. In *European Conference on Computer Vision*. Springer, 558–577.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. 2022. DiffEdit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations*.

Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. 2022. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*. Springer, 88–105.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*. Springer, 89–106.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations*.

Rinon Gal, Moab Arar, Yuval Atzmon, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. Encoder-based Domain Tuning for Fast Personalization of Text-to-Image Models. *ACM Transactions on Graphics (TOG)* 42 (2023), 1 – 13. https://api.semanticscholar.org/CorpusID:257364757

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris N. Metaxas, and Feng Yang. 2023. SVDiff: Compact Parameter Space for Diffusion Fine-Tuning. *ArXiv* abs/2303.11305 (2023).

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. 2022. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *The Eleventh International Conference on Learning Representations*.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. 2022. Imagen Video: High Definition Video Generation with Diffusion Models. *ArXiv* abs/2210.02303 (2022).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Proc. NeurIPS*.

Eliahu Horwitz and Yedid Hoshen. 2022. Conffusion: Confidence Intervals for Diffusion Models. *ArXiv* abs/2211.09795 (2022).

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Shira Iluz, Yael Vinker, Amir Hertz, Daniel Berio, Daniel Cohen-Or, and Ariel Shamir. 2023. Word-As-Image for Semantic Typography. *ACM Transactions on Graphics (TOG)* 42 (2023), 1 – 11. https://api.semanticscholar.org/CorpusID:257353586

Xuhui Jia, Yang Zhao, Kelvin C. K. Chan, Yandong Li, Han-Ying Zhang, Boqing Gong, Tingbo Hou, H. Wang, and Yu-Chuan Su. 2023. Taming Encoder for Zero Fine-tuning Image Customization with Text-to-Image Diffusion Models. *ArXiv* abs/2304.02642 (2023).

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4401–4410.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8110–8119.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6007–6017.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014).

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. arXiv:2304.02643 [cs.CV]

William H. Kruskal and Wilson Allen Wallis. 1952. Use of Ranks in One-Criterion Variance Analysis. *J. Amer. Statist. Assoc.* 47 (1952), 583–621.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1931–1941.

Gihyun Kwon and Jong Chul Ye. 2022. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18062–18071.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 9992–10002.

Elman Mansimov, Emilio Parisotto, Jimmy Ba, and Ruslan Salakhutdinov. 2016. Generating Images from Captions with Attention. *CoRR* abs/1511.02793 (2016).

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *International Conference on Learning Representations*.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6038–6047.

Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Y. Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. 2023. Dreamix: Video Diffusion Models are General Video Editors. *ArXiv* abs/2302.01329 (2023).

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. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:

245335086

Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. 2022. Mystyle: A personalized generative prior. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–10.

Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. 2023. Localizing Object-level Shape Variations with Text-to-Image Diffusion Models. *ArXiv* abs/2303.11306 (2023).

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 2065–2074. https://api.semanticscholar.org/CorpusID:232428282

Stanislav Pidhorskyi, Donald A. Adjeroh, and Gianfranco Doretto. 2020. Adversarial Latent Autoencoders. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 14092–14101.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

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).

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *Proc. ICLR*. 1060–1069.

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. 2020. Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 2287–2296.

Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. 2023. TEXTure: Text-Guided Texturing of 3D Shapes. *ACM SIGGRAPH 2023 Conference Proceedings* (2023). https://api.semanticscholar.org/CorpusID:256597953

Daniel Roich, Ron Mokady, Amit H. Bermano, and Daniel Cohen-Or. 2021. Pivotal Tuning for Latent-based Editing of Real Images. *ACM Transactions on Graphics (TOG)* 42 (2021), 1 – 13.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 10674–10685.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22500–22510.

Simo Ryu. 2022. Low-rank Adaptation for Fast Text-to-Image Diffusion Fine-tuning. https://github.com/cloneofsimo/lora.

Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. 2021a. Palette: Image-to-Image Diffusion Models. *ACM SIGGRAPH 2022 Conference Proceedings* (2021).

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.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. 2021b. Image Super-Resolution via Iterative Refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45 (2021), 4713–4726.

Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. 2022. kNN-Diffusion: Image Generation via Large-Scale Retrieval. In *The Eleventh International Conference on Learning Representations*.

Jing Shi, Wei Xiong, Zhe L. Lin, and Hyun Joon Jung. 2023. InstantBooth: Personalized Text-to-Image Generation without Test-Time Finetuning. *ArXiv* abs/2304.03411 (2023).

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2022. Make-A-Video: Text-to-Video Generation without Text-Video Data. In *The Eleventh International Conference on Learning Representations*.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*. PMLR, 2256–2265.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems* 32 (2019).

Yi-Zhe Song, Zhifei Zhang, Zhe L. Lin, Scott D. Cohen, Brian L. Price, Jianming Zhang, Soo Ye Kim, and Daniel G. Aliaga. 2022. ObjectStitch: Generative Object Compositing. *ArXiv* abs/2212.00932 (2022).

Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. 2023. Key-Locked Rank One Editing for Text-to-Image Personalization. *ACM SIGGRAPH 2023 Conference Proceedings* (2023). https://api.semanticscholar.org/CorpusID:258436985

Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. 2021. Designing an encoder for StyleGAN image manipulation. *ACM Transactions on Graphics (TOG)* 40 (2021), 1 – 14.

John W. Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics* 5 2 (1949), 99–114.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1921–1930.

Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. 2022. UniTune: Text-Driven Image Editing by Fine Tuning an Image Generation Model on a Single Image. *arXiv preprint arXiv:2210.09477* (2022).

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. 2022. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers.

Andrey Voynov, Q. Chu, Daniel Cohen-Or, and Kfir Aberman. 2023. P+: Extended Textual Conditioning in Text-to-Image Generation. *ArXiv* abs/2303.09522 (2023).

Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. 2023. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18359–18369.

Yuxiang Wei. 2023. Official Implementation of ELITE. https://github.com/csyxwei/ELITE. Accessed: 2023-05-01.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. 2023. ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation. *ArXiv* abs/2302.13848 (2023).

Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. 2021. GAN Inversion: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45 (2021), 3121–3138.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1316–1324.

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. 2023. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18381–18391.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *arXiv preprint arXiv:2206.10789* (2022).

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proc. ICCV*. 5907–5915.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2018. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence* 41, 8 (2018), 1947–1962.

Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. 2020b. In-domain gan inversion for real image editing. In *European conference on computer vision*. Springer, 592–608.

Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020a. Improved StyleGAN Embedding: Where are the Good Latents? *ArXiv* abs/2012.09036 (2020).

ACKNOWLEDGMENTS

We thank Nataniel Ruiz, Chu Qinghao, and Yael Pitch for their inspiring inputs that influenced this work. Additionally, we thank Jason Baldrige for providing valuable inputs that enhanced the quality of this project.

A ADDITIONAL EXPERIMENTS

In Appendix A.1 we start by providing additional results generated by our method. Then, in Appendix A.2 we add additional qualitative comparisons from our ablation study. Finally, in Appendix A.3 we show the results of a naïve application of TI [Gal et al. 2022] and DB [Ruiz et al. 2023] to our problem setting (multiple concepts from a single image) without the adaptation discussed in our paper.

A.1 Additional Results

In Figure 11 we provide additional results of breaking a scene into components and using them to re-synthesize novel images. Then, in Figure 12 we provide additional examples of the localized image editing application. Furthermore, in Figure 13 we provide more examples of the entangled scene decomposition application. Then, in Figure 14 we provide more examples of the image variations application. Finally, in Figure 16 and Figure 17 we provide additional qualitative comparisons of our method against the baselines.

A.2 Qualitative Ablation Study Results

As discussed in Section 4.1 in the main paper, we conducted an ablation study, which includes removing the first phase in our two-phase training scheme, removing the masked diffusion loss, removing the cross-attention loss, and removing the union-sampling. As seen in Figure 18 when removing the first training phase, the model tends to generate images that do not correspond to the target text prompt. In addition, when removing the masked diffusion loss, the model tends to learn also the background of the original image, which overrides the target text prompt. Furthermore, when removing the cross-attention loss, the model tends to mix between the concepts or replicate one of them. Finally, removing the union-sampling degrades the ability of the model to generate images with multiple concepts. In addition, increasing the probability of only one concept during the union-sampling also has a similar effect of degrading the multiple concepts generation ability.

A.3 Naïve Baselines

Existing personalization methods, such as DreamBooth (DB) [Ruiz et al. 2023] and Textual Inversion (TI) [Gal et al. 2022] take multiple images as input, rather than a single image with masks indicating the target concepts. Applying these methods to a single image without such indication results in tokens that do not necessarily correspond to the concepts that we wish to learn. In Figure 15 we provide a visual result of training TI and DB on a single image with the text prompt "a photo of $[v_1]$ and $[v_2]$ ". As expected, these approach fails to disentangle between the concepts — TI learns an arbitrary concept while DB overfits the input image.

Table 2: Personalization baselines comparison. Our method is the first to suggest a solution for the problem of *single image* with *multiple concepts* personalization. This is an extended version of Table 1 in the main paper that includes concurrent works. Only the first four methods have an open-source implementation.

Method	Single input image	Multi-concept output
Textual Inversion [Gal et al. 2022]	✗	✗
Dreambooth [Ruiz et al. 2023]	✗	✗
Custom Diffusion [Kumari et al. 2023]	✗	✓
ELITE [Wei et al. 2023]	✓	✗
E4T [Gal et al. 2023]	✓	✗
SVDiff [Han et al. 2023]	✗	✓
SuTI [Chen et al. 2023]	✗	✗
Taming [Jia et al. 2023]	✓	✗
InstantBooth [Shi et al. 2023]	✓	✗
XTI [Voynov et al. 2023]	✓	✗
Perfusion [Tewel et al. 2023]	✗	✓
Ours	✓	✓

B IMPLEMENTATION DETAILS

In the following section, we start by providing some implementation details of our method. Next, we provide more details about the automatic comparison dataset creation, as well as the automatic metrics. Finally, we provide the full details of the user study we conducted.

B.1 Method Implementation Details

We based our method, as well as the baselines (except ELITE [Wei et al. 2023]) on Stable Diffusion V2.1 [Rombach et al. 2021] implementations of the HuggingFace diffusers library [von Platen et al. 2022]. For ELITE, we used the official implementation by the authors [Wei 2023] that used Stable Diffusion V1.4, which we had to use because their encoders were trained on this model embeddings. In addition to these four baselines, many concurrent works were proposed recently, as detailed in Table 2, none of which tackles the problem of extracting *multiple concepts* from a *single image*.

As explained in Section 3 of the main paper, our method is divided into two stages: in the first stage we optimize only the text embeddings with a high learning rate of $5e^{-4}$, while in the second stage, we train both the UNet weights, and the text encoder weights with a small learning rate of $2e^{-6}$. For both stages, we used Adam optimizer [Kingma and Ba 2014] with $\beta_1 = 0.9, \beta_2 = 0.99$ and a weight decay of $1e^{-8}$. For all of our experiments, we used 400 training steps for each one of the stages, which we found to work well empirically. When applying the masked version of the baselines, we used the corresponding learning rate and optimized parameters as our method. We performed the union-sampling in both of the training stages.

The UNet of the Stable Diffusion models consisted of a series of self-attention layers followed by cross-attention layers that inject the textual information into the image formation process. This is done in various resolutions of 8, 16, 32, 64. As was shown in [Hertz et al. 2022], these cross-attention layers also control the layout of

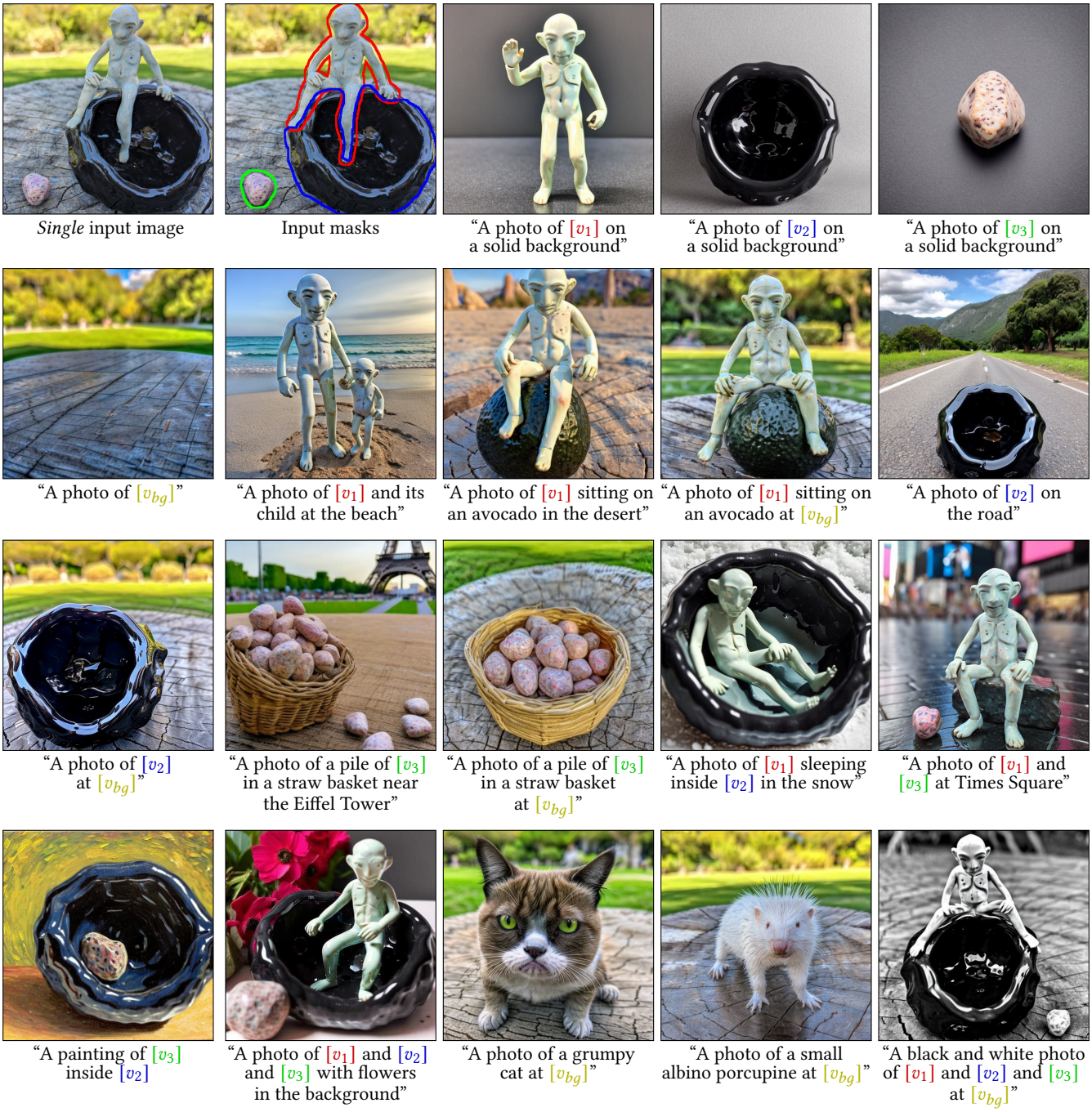

Figure 11: Additional break-a-scene results: a scene decomposed into 3 parts and a background, which are then re-synthesized in different contexts and combinations.

the generated scene, and can be utilized for generating images with the same structure but with different semantics, or edit generated images. As explained in Section 3 of the main paper, we utilize these cross-attention maps for disentangling between the learned concepts. To this end, we average all the cross-attention maps corresponding to each one of the newly-added personalized tokens at resolution 16×16, which was shown by [Hertz et al. 2022] to contains most of the semantics, and normalized them to range $[0, 1]$. For brevity, we refer to this normalized averages cross-attention map as $CA_\theta(v_i, z_t)$, the cross-attention map between the token v_i and the noisy latent z_t.

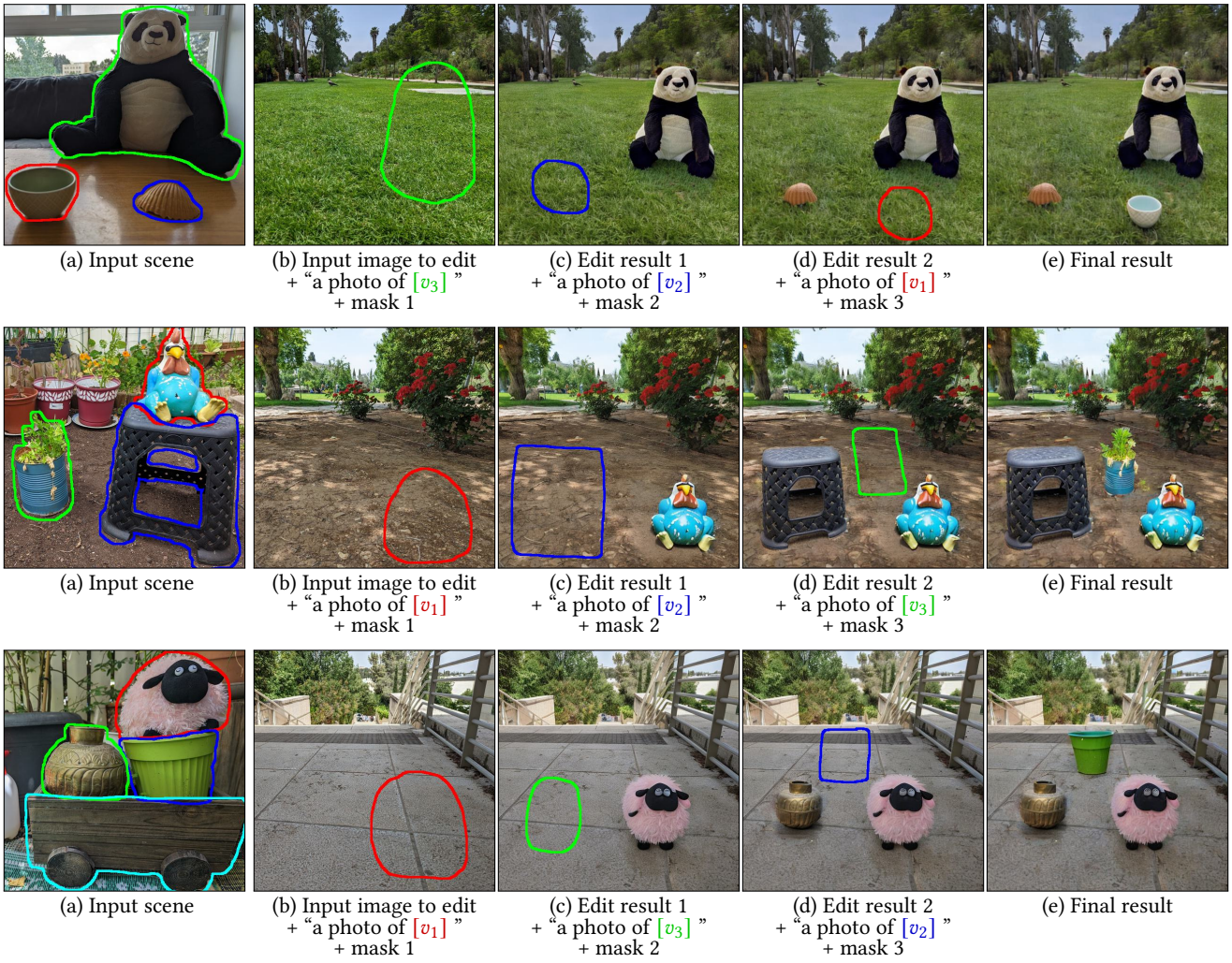

Figure 12: Additional examples of local image editing: given an input scene (a), we extract the indicated concepts using our method. Given an additional input image to edit (b) along with a mask indicating the edit area, and a guiding text prompt, we use Blended Latent Diffusion [Avrahami et al. 2023a, 2022] to obtain the first edit result (c). The process (provide mask and prompt, apply Blended Latent Diffusion) can be repeated (c–d), until the final outcome is obtained (e).

B.2 Automatic Dataset Creation

As explained in Section 4.1 in the main paper, we created an automated pipeline for creating a comparisons dataset and use it to compare our method (quantitatively and via a user study). To this end, we use COCO [Lin et al. 2014] dataset, which contains images along with their instance segmentation masks. We crop COCO images into a square shape, and filter only those that contain at least two segments of distinct "things" type, with each segment occupying at least 15% of the image. We also filter out concepts from COCO classes that are hard to distinguish from each other (orange, banana, broccoli, carrot, zebra, giraffe). Using this method, we extracted 50 scenes of different types. Next, we paired each of these inputs with a text prompt from a fixed list, e.g., "a photo of {tokens} in the snow", where {tokens} was iterated on all the combinations of the powerset of the input tokens, yielding a total number of 5400 generations per baseline. Figure 17 presents a qualitative comparison of the baselines against our method on this automatically generated dataset.

The fixed formats that we used are:

- "a photo of {tokens} at the beach"
- "a photo of {tokens} in the jungle"
- "a photo of {tokens} in the snow"
- "a photo of {tokens} in the street"
- "a photo of {tokens} on top of a pink fabric"
- "a photo of {tokens} on top of a wooden floor"
- "a photo of {tokens} with a city in the background"
- "a photo of {tokens} with a mountain in the background"
- "a photo of {tokens} with the Eiffel tower in the background"
- "a photo of {tokens} floating on top of water"

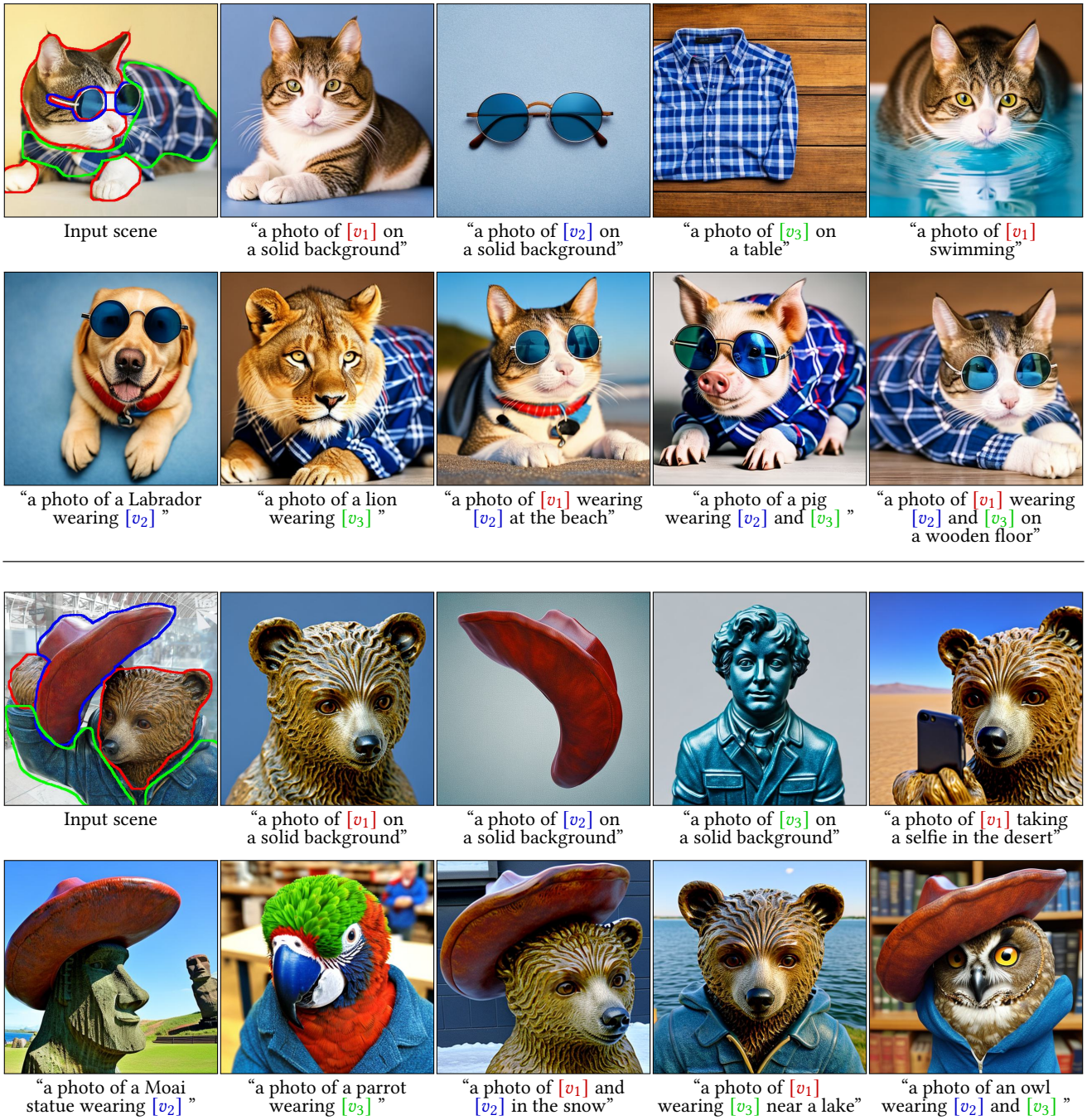

Figure 13: Additional examples of entangled scene decomposition: given a single input image of several spatially-entangled concepts, our method is able to disentangle them and generate novel images with them, separately or jointly.

As explained in Section 4.1 in the main paper, we focused on two evaluation metrics: prompt similarity and identity similarity. For calculating the prompt similarity we used CLIP [Radford et al. 2021] model ViT-L/14 [Dosovitskiy et al. 2020] implementation by HuggingFace and calculated normalized cosine similarity of the CLIP text embeddings of the input prompt (the tokens were replaced with the ground-truth classes) and CLIP image embedding of the generated image. For calculating the identity similarity, we offered a metric that supports the multi-subject case: for each generation we compare the masked version of the input image (by the input mask from the COCO dataset) with a masked version of

Figure 14: Additional examples of image variations applications: given a single input image of several concepts, our method is able to generate many variations of the image. Credits: pxhere

the generated image, which we acquire by utilizing a pre-trained image segmentation model MaskFormer [Cheng et al. 2021] that was trained on COCO panoptic segmentation (large-sized version, SWIN [Liu et al. 2021] backbone) implemented by HuggingFace. For the image embeddings comparison, we used DINO model [Caron et al. 2021] (base-sized model, patch size 16) that was shown [Ruiz et al. 2023] to better encompass the object's identity.

B.3 User Study Details

As described in Section 4.1 in the main paper, we conducted a user study employing the Amazon Mechanical Turk (AMT) in order to assess the human perception of the metrics of interest: prompt similarity and identity similarity. For assessing the prompt correspondence, we instructed the workers "For each of the following images, please rank on a scale of 1 to 5 its correspondence to this text description: {PROMPT}" where {PROMPT} is the modified text prompt resulted by replacing the special token with the class textual token (e.g., "a photo of a cat at the beach" instead of "a photo of [v_1] at the beach" which was used to create the image). All the baselines, as well as our method, were presented in the same page, and the evaluators rated each result by a slider from 1 ("Do not match at all") to 5 ("Match perfectly "). For assessing identity similarity, we showed a masked version of the input image that contains only the object being generated, put it next to each one of the baseline results, and instructed the workers "For each of the following image pairs, please rank on a scale of 1 to 5 if they contain the same object (1 means that they contain totally different objects and 5 means that they contain exactly the same object). The images can have different backgrounds". The questions were presented to the raters in a random order, and we collected three ratings per question, resulting in 1215 ratings per task (prompt similarity/identity similarity). The time allotted per image-pair task was one hour, to allow the raters to properly evaluate the results without time pressure.

We conducted a statistical analysis of our user study by validating that the difference between all the conditions is statistically significant using Kruskal-Wallis [Kruskal and Wallis 1952] test ($p < 10^{-213}$). In addition, we used Tukey's honestly significant difference procedure [Tukey 1949] to show that the comparison of

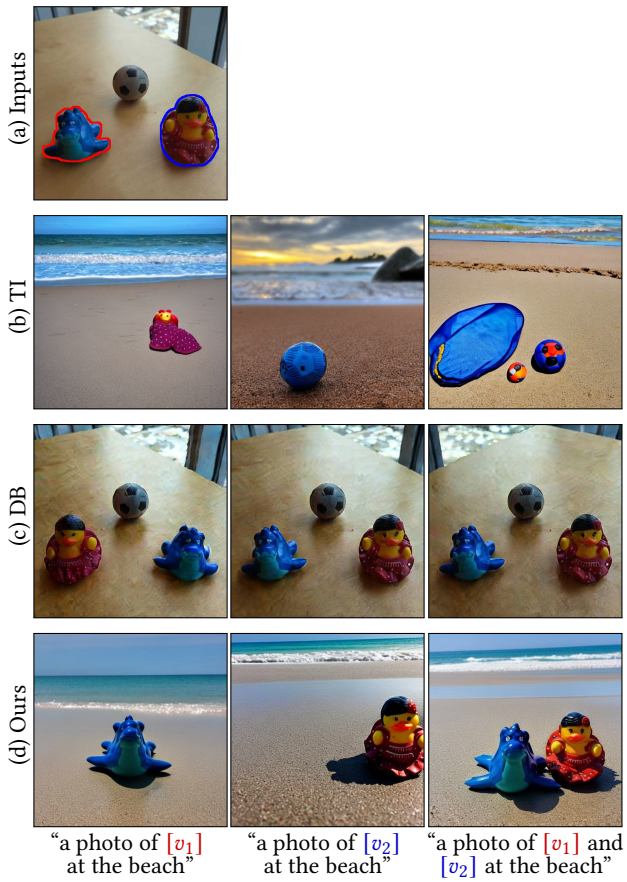

Figure 15: Naive TI and DB adaptations: given an (a) input scene, trying naïvely running (b) TI and (c) DB on the input image. As expected, these approach fails to disentangle between the concepts — TI learns an arbitrary concept while DB overfits the input image. On the other hand, (d) our method is able to learn the identity of the concepts while taking into account the text prompt.

Table 3: Statistical analysis. We use Tukey's honestly significant difference procedure [Tukey 1949] to test whether the differences between mean scores in our user study are statistically significant.

Method 1	**Method 2**	Prompt similarity p-value	Identities similarity p-value
TI-m	Ours	$p < 10^{-10}$	$p < 10^{-10}$
DB-m	Ours	$p < 0.05$	$p < 10^{-10}$
CD-m	Ours	$p < 10^{-7}$	$p < 10^{-10}$
ELITE	Ours	$p < 10^{-10}$	$p < 10^{-10}$

our method against all the baselines is statistically significant, as detailed in Table 3. The means and variances of the user study are reported in Table 4.

Table 4: Users' rankings means and variances. the means and variances of the rankings that are reported in the user study.

Method	Identity similarity	Prompt similarity
TI-m	2.69 ± 1.3	3.88 ± 1.21
DB-m	3.97 ± 0.95	2.37 ± 1.11
CD-m	2.47 ± 1.3	4.08 ± 1.12
ELITE	3.05 ± 1.31	3.53 ± 1.31
Ours	3.56 ± 1.27	3.85 ± 1.21

B.4 Blended Latent Diffusion Integration

As explained in Section 4.2 in the main paper, in order to edit an image using the extracted concepts from another image, we utilized Blended Latent Diffusion [Avrahami et al. 2023a, 2022] off-the-shelf text-driven image editing method. As shown in Figure 12, we can perform it in an iterative manner, editing the image region-by-region. That way, we can edit the image in an elaborated manner by giving a different text prompt per region.

C SOCIETAL IMPACT

Our method may help democratizing content creation, empowering individuals with limited artistic skills or resources to produce visually engaging content. This may not only open up opportunities for individuals who were previously excluded, but also foster a more diverse and inclusive creative landscape.

In addition, it can help generate visuals that align with specific rare cultural contexts, where the input may be scarce and contain a single image. This may enhance cultural appreciation, foster a sense of belonging, and promote intercultural understanding.

On the other hand, our method, may cause intellectual property and copyright issues when being used on an existing copyrighted content as reference. In addition, malicious users can exploit this model to create realistic but fabricated images, potentially deceiving other individuals.

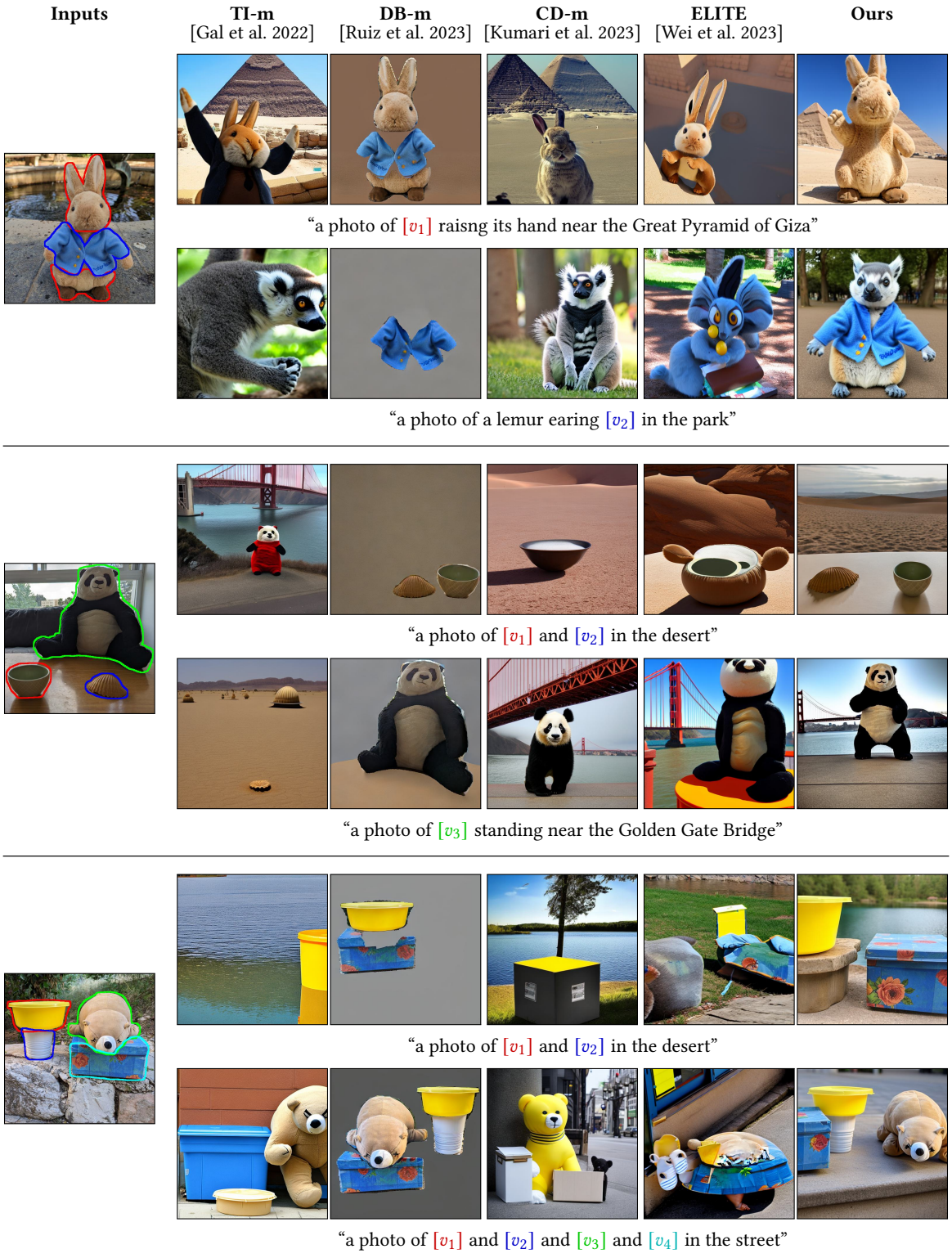

Figure 16: A qualitative comparison between several baselines and our method. As can be seen, TI-m and CD-m struggle with preserving the concept identities, while the images generated by DB-m effectively ignore the text prompt. ELITE preserves the identities better than TI-m/CD-m, but the concepts are still not recognizable enough, especially when more than one concept is generated. Finally, our method is able to preserve the identities as well as to follow the text prompt, even when learning four different concepts (bottom row).

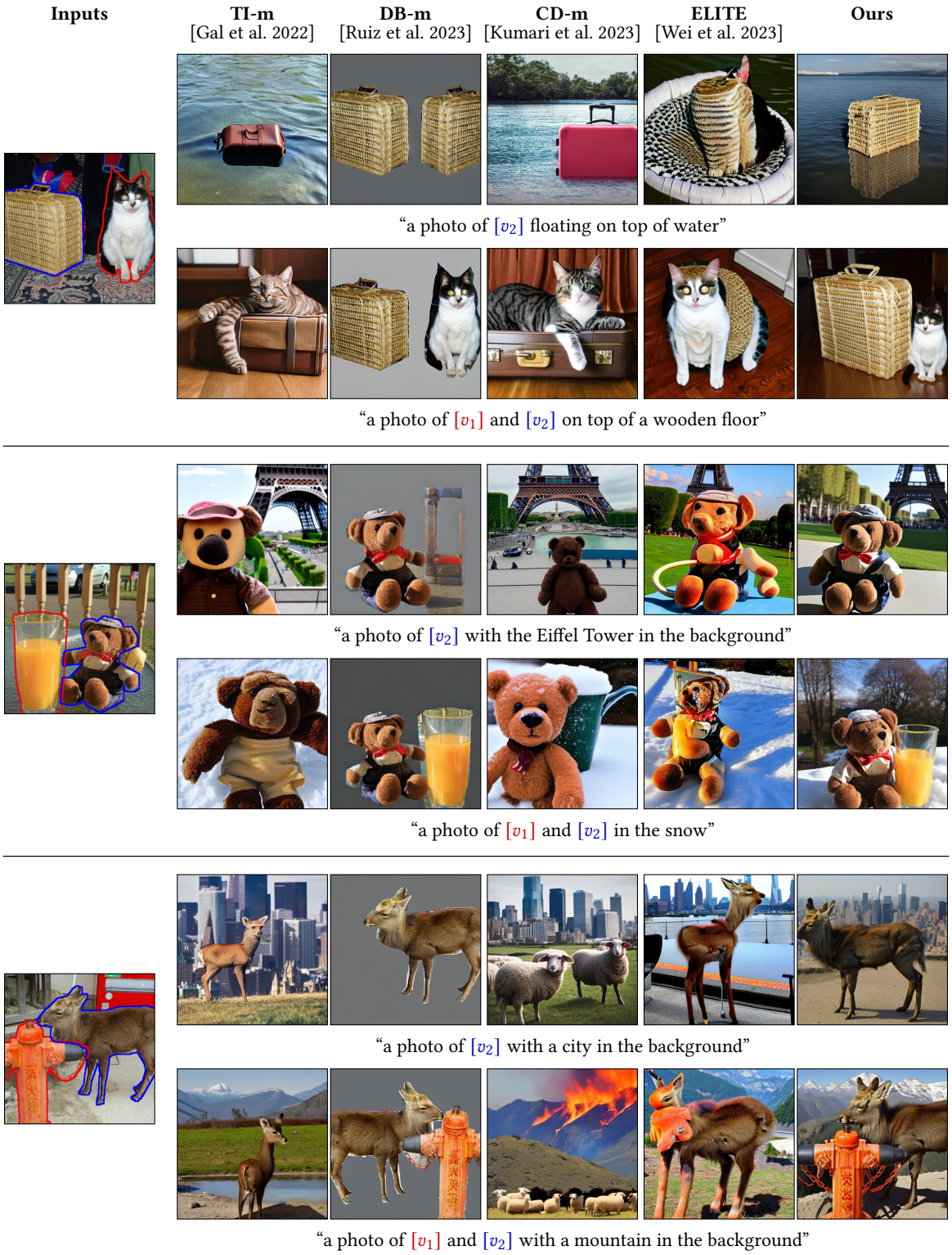

Figure 17: Automatic dataset qualitative comparison: we compare our method qualitatively against the baselines on the dataset that was generated automatically, as explained in Appendix B.2. As we can see, TI-m and CD-m struggle with preserving the concept identities, while DB-m struggle with generating an image the corresponds to the text prompt. ELITE is better preserving the identities than TI-m/CD-m, but they are still not recognizable enough, especially when trying to generate more than one concept. Finally, our method is able to preserve the identities as well as correspond to the text prompt, and even support generating up to four different concepts.

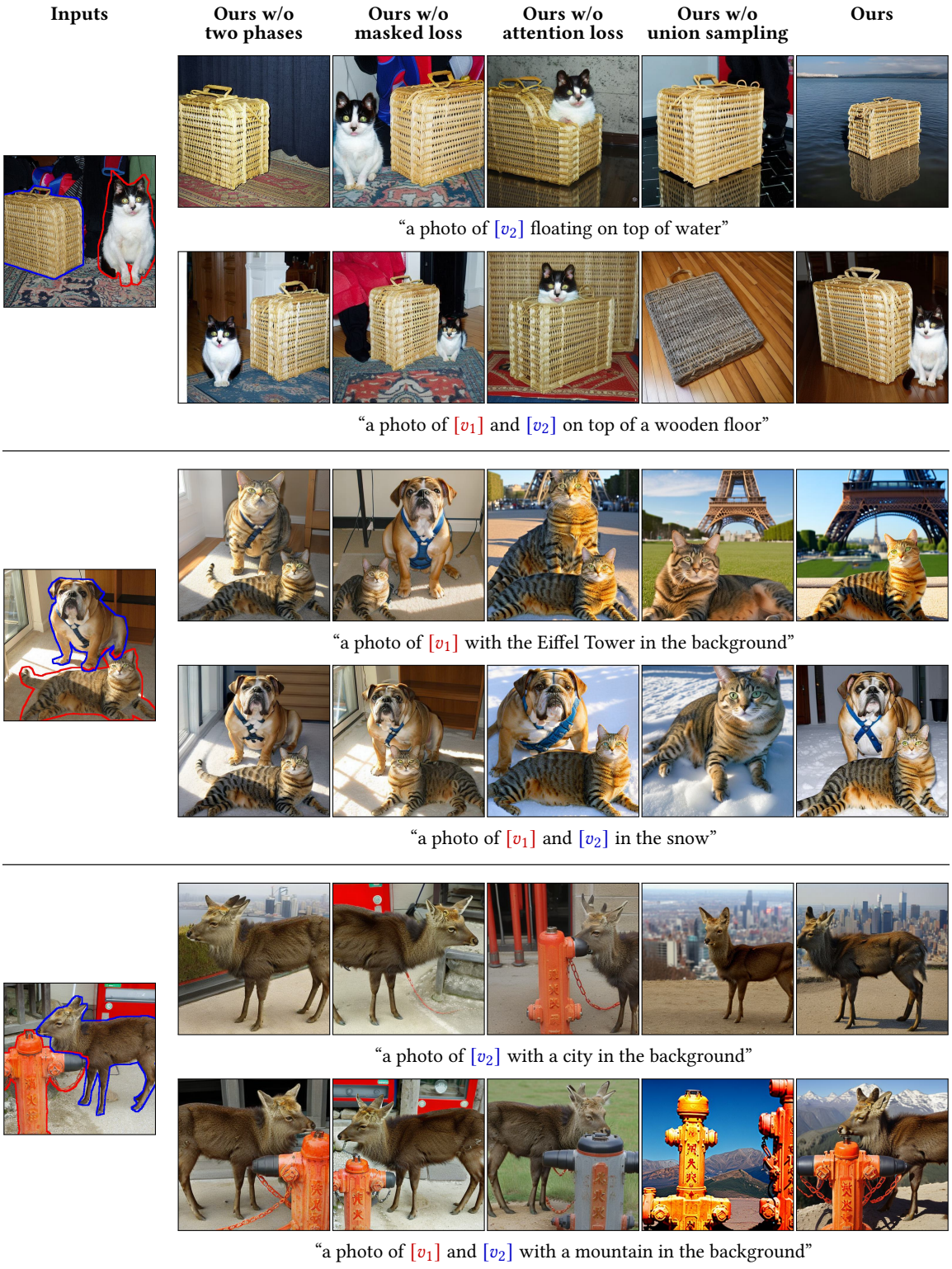

Figure 18: Qualitative ablation study: we conduct an ablation study by removing the first phase in our two-phase training regime, removing the masked diffusion loss, removing the cross-attention loss, and removing the union-sampling. As can be seen, when removing the first training phase, the model tends to correspond less to the input text prompt. In addition, when removing the masked loss, the model tends to learn also the background, which diminishes the target prompt correspondence. Furthermore, when removing the cross-attention loss, the model tends to mix between the concepts or replicate one of them. Finally, when removing the union-sampling, the model struggles with generating images with multiple concepts.