Attention Calibration for Disentangled Text-to-Image Personalization

Yanbing Zhang[1,2], Mengping Yang[1,2], Qin Zhou[1,2*], Zhe Wang[1,2*]
[1] Department of Computer Science and Engineering, ECUST, China
[2] Key Laboratory of Smart Manufacturing in Energy Chemical Process, ECUST, China
{zhangyanbing, mengpingyang}@mail.ecust.edu.cn, {sunniezq, wangzhe}@ecust.edu.cn

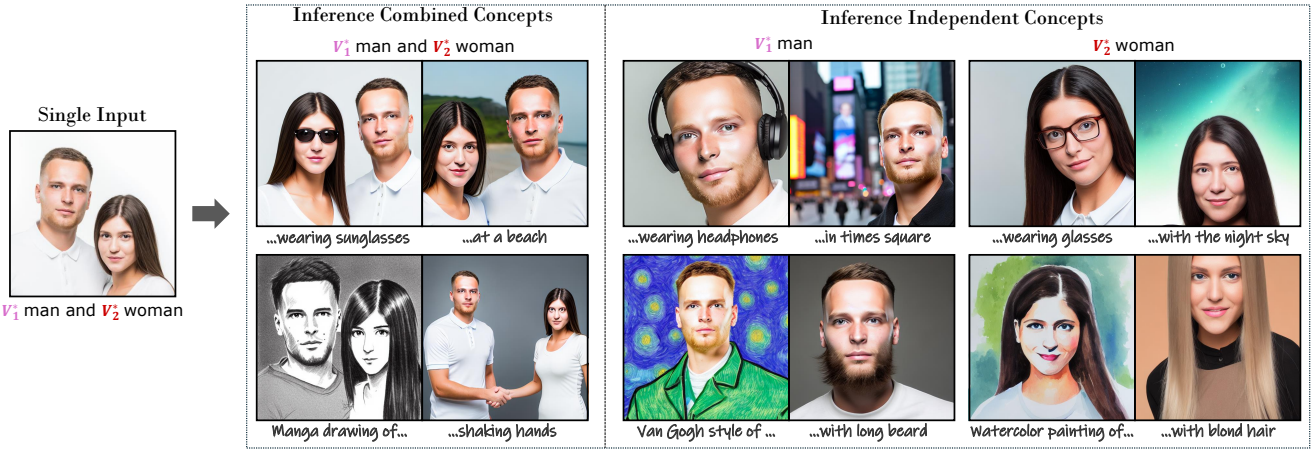

Figure 1. Given one *individual* image from specific users, our proposed method is capable of producing *customized* images for each concept contained in the input image, *e.g.,* given a single input image with a man and a woman, our method excels in achieving innovative renditions of both combined (*left*) and independent (*right*) concepts, without compromising the fidelity and identity preservation, and more importantly, manifesting satisfactory interactive generation conditioned by various text prompts. Note that we employ notation V_i^* to denote the modifier of the i-th concept. Our code and data will be publicly available at: https://github.com/Monalissaa/DisenDiff.

Abstract

Recent thrilling progress in large-scale text-to-image (T2I) models has unlocked unprecedented synthesis quality of AI-generated content (AIGC) including image generation, 3D and video composition. Further, personalized techniques enable appealing customized production of a novel concept given only several images as reference. However, an intriguing problem persists: Is it possible to capture ***multiple, novel concepts*** *from* ***one single reference image****? In this paper, we identify that existing approaches fail to preserve visual consistency with the reference image and eliminate cross-influence from concepts. To alleviate this, we propose an attention calibration mechanism to improve the concept-level understanding of the T2I model. Specifically, we first introduce new learnable modifiers bound with classes to capture attributes of multiple concepts. Then, the classes are separated and strengthened following the activation of the cross-attention operation, ensuring comprehensive and self-contained concepts. Additionally, we suppress the attention activation of different classes to mitigate mutual influence among concepts. Together, our proposed method, dubbed* ***DisenDiff****, can learn disentangled multiple concepts from one single image and produce novel customized images with learned concepts. We demonstrate that our method outperforms the current state of the art in both qualitative and quantitative evaluations. More importantly, our proposed techniques are compatible with LoRA and inpainting pipelines, enabling more interactive experiences.*

1. Introduction

Recently developed large-scale text-to-image models [1, 37, 39, 41] have shown unprecedented capabilities in synthesizing high-quality and diverse images based on a target text prompt. Built on these models, personalized techniques [11, 40] are further introduced to customize the models for

*Corresponding author

Figure 2. **Failure case of Custom Diffusion [23].** In the third column, we show the example encompassing two failure settings: appearance inconsistency with the input image and ambiguous object not included in the target text. In the second column, we show the result from our method.

synthesizing personal concepts with sufficient fidelity.

Given as input just a few images of the personal concepts (e.g., family, friends, pets, or individual objects), personalized text-to-image models aim to learn a new word embedding to represent a specific concept [45, 50]. However, existing methods still lack the flexibility to render all existing concepts in a given image, or only focus on a specific concept [12, 25]. Given a unique photo from a user (which could be people rarely seen together or uncommon furniture pieces), with multiple concepts occurring in the complex scene, the user naturally desires the ability to freely synthesize the concepts by composing multiple objects or focusing on only one of them. For example, two specific individuals at a beach, or alternatively, one of them in Times Square, as shown in Fig. 1.

To achieve flexible renditions of the concepts, instead of using a single new word to represent one concept [19, 23], we employ multiple new words to represent multiple concepts. For example, considering an image containing a distinct chair and lamp (as shown in Fig. 2), we utilize the prompt "V_1^* chair and V_2^* lamp" to distinguish between them, with "V_1^*" serving as the modifier for "chair" and "V_2^*" as the modifier for "lamp". This intuitive formulation poses two key challenges. Firstly, the new word embeddings are likely to map confusing information, failing to maintain visual-fidelity to the target concepts. Secondly, with a relatively small training set (e.g., only one image), the model is prone to synthesizing multiple subjects, even when the target prompt pertains to a single concept. For example, as depicted in Fig. 2, the ideal output should exclusively feature the specified lamp when the target text is "A V_2^* lamp". Nonetheless, the image generated by the current state-of-the-art model not only includes a lamp that doesn't match the color and texture of the input image but also involves a chair that shouldn't be present.

In this paper, we propose a novel personalized T2I model, referred to as *DisenDiff* (i.e., Disentangled Diffusion), to address the above-mentioned issues. To preserve the good generalization ability in pre-trained large-scale models, we follow [23, 45] to only update the light-weight modules (W_K and W_V matrices) within the cross-attention units along with new token embeddings to extend concepts. Our key insight is that current methods lack the necessary guidance for the optimization process, resulting in cluttered attention maps (as shown in Fig. 4, the first row). Consequently, existing methods struggle to synthesize each concept effectively.

Based on the above observations, we strive to generate precise attention maps from the following two aspects. Building on the discovery that the attention map of the class token can roughly align with the location of the concept, then we propose a modifier-class alignment term to bind the attention map of each new modifier with its corresponding class token, correcting attention to focus on the region of the related concept. However, the attention maps of different class tokens often exhibit overlaps, leading to the incorrect attribute binding [4] and mutual entanglement. To achieve effective decoupling, we introduce the separate and strengthen (s&s) strategy to allow flexibly synthesizing each concept independently. By minimizing the overlapping regions between the attention maps of different class tokens, we can effectively mitigate the co-occurring issue when targeting at a specific concept. To further enhance the independence of concepts, we introduce a suppression technique to sharpen the boundaries of class tokens' attention maps. Our contributions are summarized below:

- We propose *DisenDiff* to comprehend multiple personal concepts from only a single image. By using diverse target texts, it can render combined/independent concepts in imaginary contexts while preserving high fidelity to the input image.
- We employ two key constraints to attain precise attention maps for crucial tokens. The binding constraint locates new modifiers to different concepts, while the s&s constraint decouples these concepts.
- We conduct experiments on various datasets and demonstrate that our method outperforms the current state of the art in quantitative and qualitative aspects. Additionally, we show the flexibility of our approach by applying it to extended tasks.

2. Related Work

Text-to-image generative models. The objective of text-to-image (T2I) tasks [29, 56] is to generate an image corresponding to a given textual description. Thanks to large-scale datasets [3, 42] and advancements in language models [21, 34, 35], T2I models have witnessed remarkable progress. While Generative adversarial networks (GANs) [20, 27, 38, 53] and autoregressive (AR) transformers [8, 10, 36, 52] have delivered impressive results, diffusion models [7, 16] have taken the lead in T2I generation. These models employ denoising processes in image space [1, 17, 31, 41, 49] or latent space [13, 37, 39], resulting in

unprecedented image generation quality. However, they encounter challenges when generating specific objects, such as custom furniture, even with detailed prompts. We aim to augment these models to accurately capture the appearances of novel concepts from real-world images.

Text-guided image editing. With the surge of powerful T2I models, numerous studies have delved into enhancing the controllability of diffusion models to cater to diverse user demands. Approaches such as [4, 9, 47] refine the cross-attention units to encompass all subject tokens, motivating the model to fully convey the semantics in the input prompt. Techniques like [5, 26, 32] implement region control in T2I generation by using bounding boxes and paired object labels as inputs. Additionally, [54] and [48] harness pre-trained diffusion models for image-to-image translation. A substantial body of work also focuses on local or global modifications of single images using existing T2I models. Notable examples include SINE [55] and UniTune [46], which achieve image editing by fine-tuning the diffusion model. Other methods like prompt-to-prompt [14], null-text inversion [30], and [33] impose constraints on latent noise during inference time without model training. While our objectives share some common ground with these methods, our primary focus is optimizing the model to seamlessly extend personalized concepts into new prompts.

T2I personalization Personalization techniques adapt diffusion models to learn new concepts from user-provided images, often relying on a small dataset of 3-5 images or even a single image. Textual Inversion [11] uses pseudo-words to represent new concepts through a visual reconstruction objective. To leverage semantic priors from pre-trained models, DreamBooth [40] utilizes a unique identifier and class name within the input text to represent new concepts. Custom-Diffusion [23] and Perfusion [45] compose multiple new concepts by updating only the cross-attention Keys and Values along with new token embeddings. When working with a dataset containing just a single image, current methods [12, 19, 25, 50] typically begin with additional domain-specific pre-training on a large dataset before adapting to the new concept. In contrast to these methods, we aim to address the more challenging problem of acquiring multiple concepts from a single image without domain-specific pre-training.

3. Method

Our objective is to understand multiple concepts within a single image. To this end, we propose a novel attention calibration mechanism to help generate accurate cross-attention maps in our T2I model. Firstly, the cross-attention maps are calculated as the activation responses between each word of the input text and the intermediate visual features. Then, we impose constraints on the cross-attention maps between both the modifier-class token pairs and class-class token pairs to bind the cross-attention maps of each modifier with its corresponding class (modifier-class constraint), as well as to ensure full comprehension of each class and separation between different classes (class-class constraint). To further mitigate the cross-interference issue in our T2I model, we introduce a suppression technique to obtain a sharper attention map for each class token. A schematic workflow of our method is presented in Fig. 3.

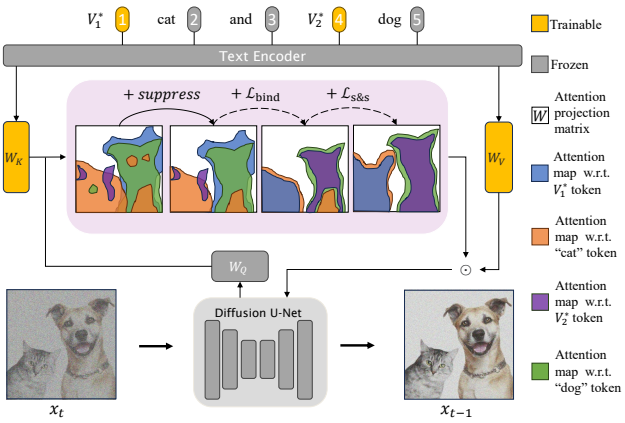

Figure 3. **Method overview.** Our method applies constraints to the cross-attention maps of crucial tokens, ensuring the accurate representation of multiple concepts. We introduce new modifiers, denoted as V_i^*, along with the i-th class name, to represent the i-th personalized concept. Our attention calibration mechanism mainly includes three parts: the suppression technique performs self-sharpening and filters noisy small patches, the $\mathcal{L}_{\text{bind}}$ loss steers new modifiers towards the corresponding classes, and the $\mathcal{L}_{\text{s\&s}}$ loss guarantees the independence and completeness of the learned concepts.

3.1. Preliminary

Stable Diffusion. In our experiments, we use Stable Diffusion [43] as our backbone model, inheriting the structure of the Latent Diffusion Model (LDM) [39]. It primarily consists of three components: a pre-trained text encoder τ_θ from CLIP [34], a VAE [22] model $\mathcal{E}$, and a U-Net diffusion model ϵ_θ trained on the latent space z of the pre-trained VAE. Given the noisy latent code z_t at t timestep, the diffusion model predicts the random added noise ϵ. The training objective of the diffusion model is formulated as follows:

$$\mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\left\|\epsilon-\epsilon_\theta\left(z_t,t,\tau_\theta(y)\right)\right\|_2^2\right], \quad (1)$$

where x denotes the input image, y is the input text. Following [39], prior knowledge in CLIP is integrated via the cross-attention mechanism.

Integrating textual features via cross-attention. Formally, the intermediate spatial representation $\phi(z_t)$ of the denoiser U-Net is mapped to a query matrix $Q = W_Q \cdot \phi(z_t)$, while text embeddings $\tau_\theta(y)$ are mapped to a key matrix $K = W_K \cdot \tau_\theta(y)$ and a value matrix $V = W_V \cdot \tau_\theta(y)$, us-

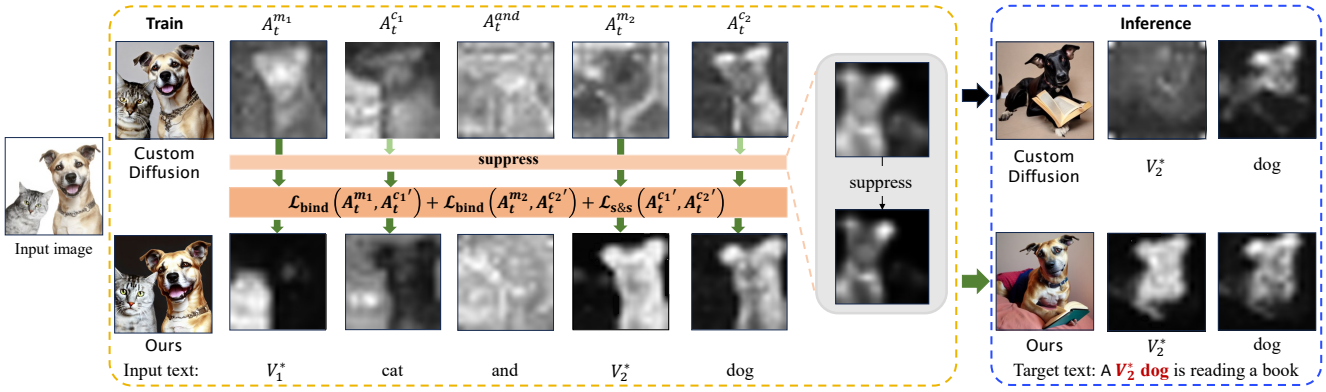

Figure 4. **Comparison of generated attention maps and images.** The first row displays the results of Custom Diffusion [23], while the second row shows our results. During the training stage, when we obtain accurate attention maps for important tokens (left), it leads to the ideal output during the inference stage (right), maintaining high-concept similarity with the input image.

ing learnable projection matrices W_Q, W_K, and W_V. Then, the cross-attention maps are obtained as:

$$A_t = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right), \tag{2}$$

where d is the projection dimension of keys K and queries Q. Here, $A_t \in \mathbb{R}^{r\times r\times N}$, r is the spatial dimension of the $\phi(z_t)$, and N is the number of input tokens. The updated spatial representations integrating text priors are then obtained as $\phi(z_t) = A_t V$, as illustrated in Fig. 3.

Text encoding. Generally, during training of a T2I system, a suitable text prompt is required in addition to the selected single image. In this paper, we adopt a manner similar to [40], incorporating new modifiers and the classes to be modified into the input text. For example, if the target image contains a cat and a dog, the text prompt would be "V_1^* cat and V_2^* dog". The modifier tokens "V_i^*" are initialized with rare vocabulary. Given only a single training image, the T2I model will likely lack the diversity of generation, known as the language drift [24, 28] problem. Using our text prompt, we can easily select regularized images with the same caption to mitigate the issue of language drift, enabling our model to generate a variety of cats and dogs (not limited to the ones present in the target images, as shown in Fig. 5, left of the second row).

Current methods are prone to overfitting when the training data only consists of a single image, resulting in ambiguous attention maps for each token (as shown in the first row of Fig. 4). As demonstrated in P2P [14], the spatial layout and geometry of the generated images depend on the cross-attention maps. Therefore, our primary focus is to optimize the model to produce accurate cross-attention maps, elaborated in the following part.

3.2. Coherent binding of modifiers with classes

Based on the cross-attention maps (A_t) obtained by a previous method (shown in Fig. 4, the first row), we can observe that while A_t of new modifiers are chaotic ($A_t^{m_1}$ and $A_t^{m_2}$), cross-attention of class tokens can roughly capture the semantic boundaries ($A_t^{c_1}$ and $A_t^{c_2}$). We attribute it to the fact that the majority of parameters in the T2I models are frozen, preserving the category information of class tokens. To aid the new modifiers in understanding their responsibilities, we define the constraint to bind the cross-attention maps of modifiers with their corresponding class tokens as

$$\mathcal{L}_{\text{bind}}\left(A_t^{m_i}, A_t^{c_i}\right) = 1 - \frac{A_t^{m_i} \cap A_t^{c_i}}{A_t^{m_i} \cup A_t^{c_i}}, \tag{3}$$

where $A_t^{m_i}$ and $A_t^{c_i}$ represent the attention map of the i-th modifier and the i-th class at t timestep, respectively. The $\mathcal{L}_{\text{bind}}$ loss is formulated to reduce the intersection over union (IoU) [51] between these two attention maps, encouraging a close alignment between the activations of the modifiers and the class tokens. To prevent substantial influence on $A_t^{c_i}$, we detach its gradient during the loss computation.

Nonetheless, there are two potential issues when we directly apply this constraint. Given that A_t is the result of the Softmax operation (i.e., $\sum_{i=1}^{N} A_t^i(h,w) = 1$, where $A_t^i(h,w)$ denotes the activation of the i-th token at pixel (h,w)), input tokens would contend for attention at the same position. Consequently, a precise pixel-to-pixel correspondence between $A_t^{m_i}$ and $A_t^{c_i}$ can not be established. Furthermore, our intention is for the activations of $A_t^{m_i}$ to fully encompass the corresponding object, thereby capturing all its attributes comprehensively. However, as depicted in Fig. 4, it is evident that within the object region, certain activations of $A_t^{c_2}$ exhibit high values, while others appear considerably lower. This poses a challenge for the attention $A_t^{m_i}$ to sustain a comprehensive focus on the object. To address these challenges, we employ a Gaussian filter on A_t, which leads to the generation of smooth attention maps referred to as $G(A_t)$. This smoothing process helps to alleviate the pixel-wise competition among tokens and facilitates more comprehensive attention to the object. Consequently,

by using the loss function $L_{\text{bind}}\left(G(A_t^{m_i}), G(A_t^{c_i})\right)$, we encourage $A_t^{m_i}$ to have coherent attention areas with $A_t^{c_i}$, while achieving a broader coverage of the object, without the need for precise point-to-point binding. For simplicity, in the subsequent sections of this paper, unless explicitly specified otherwise, we apply a Gaussian filter to A_t.

3.3. Separating and strengthening attention maps for multiple classes

Given a single image as the training set, it's inevitable for one class token to attend to multiple concepts simultaneously. For instance, in the first row of Fig. 4, specifically in $A_t^{c_1}$, the attention dedicated to the "cat" token is not solely limited to the "cat" concept. It also exhibits some degree of attention towards the "dog" concept. Thus, $A_t^{m_1}$ incorporates attributes associated with the "dog" concept due to its binding with $A_t^{c_1}$. To ensure independent editing of concepts without interference, it is necessary to separate the attention regions of different objects (i.e., $A_t^{c_i}$ and $A_t^{c_j}$). A straightforward approach is to minimize the overlap between attention maps of different object tokens as

$$\mathcal{L}_{\text{separate}}\left(A_t^{c_i}, A_t^{c_j}\right) = A_t^{c_i} \cap A_t^{c_j}. \tag{4}$$

The utilization of $\mathcal{L}_{\text{separate}}$ effectively prevents the activations of class tokens from overlapping. However, it may come with a side effect of reducing the area of $A_t^{c_i}$, potentially leading to a loss of identity for the corresponding class, which can be found in the supplement. To simultaneously minimize the overlap among attention maps and preserve the class identity, we design the following constraint,

$$\mathcal{L}_{\text{s\&s}}\left(A_t^{c_i}, A_t^{c_j}\right) = \frac{A_t^{c_i} \cap A_t^{c_j}}{A_t^{c_i} \cup A_t^{c_j}}, \tag{5}$$

where "s&s" stands for "separate and strengthen." The $\mathcal{L}_{\text{s\&s}}$ loss strikes a balance between avoiding overlap with other objects and ensuring comprehensive coverage of the target object, thus improving the accuracy and fidelity of the attention mechanism.

Suppression. The utilization of the $\mathcal{L}_{\text{s\&s}}$ loss can potentially lead to another issue where the attention map $A_t^{c_1}$ captures a significant portion of the activations, while $A_t^{c_2}$ exhibits very few activations. This imbalance in activation distribution between different class tokens can result in an uneven emphasis on certain classes. To address it, we introduce a suppression mechanism. Specifically, before computing the $\mathcal{L}_{\text{s\&s}}$, we apply an element-wise multiplication operation to $A_t^{c_i}$ (i.e., $f_m(A_t^{c_i}) = A_t^{c_i} \odot A_t^{c_i}$). Given that activations fall within the range of $[0, 1]$, $f_m(A_t^{c_i})$ filters out activations that are less important for the class. As a result, the loss $L_{\text{s\&s}}(f_m(A_t^{c_i}), f_m(A_t^{c_j}))$ is designed to separate and strengthen their attentions, preventing encroachment upon other classes from within its own boundaries. Additionally, $A_t^{m_i}$ can be bound with a more distinct $A_t^{c_i}$.

In summary, the total training loss is formulated as:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{base}} &+ \sum_{i=1}^{\mathcal{S}} \mathcal{L}_{\text{bind}}\left(G(A_t^{m_i}), f_m(G(A_t^{c_i}))\right) \\ &+ \sum_{i=1}^{\mathcal{S}} \sum_{j=i+1}^{\mathcal{S}} \mathcal{L}_{\text{s\&s}}\left(f_m(G(A_t^{c_i})), f_m(G(A_t^{c_j}))\right),\end{aligned} \tag{6}$$

where $\mathcal{S}$ is the number of classes in the input image, and $\mathcal{L}_{\text{base}}$ is the base loss of the T2I model in Eq. (1). $\mathcal{L}_{\text{s\&s}}$ is responsible for refining the attention maps related to class tokens, while $\mathcal{L}_{\text{bind}}$ is responsible for constraining new modifier tokens to acquire correct attributes. The auxiliary functions $G(\cdot)$ and $f_m(\cdot)$ facilitate the optimization process. The synergy among these constraints results in the generation of precise and interpretable attention maps for input tokens, shown in the second row of Fig. 4.

4. Experiments

4.1. Experimental Settings

Datasets. We conducted experiments on ten datasets spanning a large range of categories including people, animals, furniture, and people with pets/toys. Please note, instead of concentrating only on one concept, our datasets contain *two distinct concepts* within each image. During the inference phase, we test 30 different prompts for each image: 10 for combined concepts, 10 specifically targeting the first concept, and 10 focusing on the second concept.

Compared methods. We compare with three personalized T2I methods, which all utilize new word embeddings to represent novel concepts. (1) Textual Inversion (TI): In TI, only the new token embedding representing the novel concept is updated, while the other parameters remain frozen. (2) DreamBooth (DB): DB updates all layers of the T2I model to maintain visual fidelity and employs a prior preservation loss to mitigate language drift. (3) Custom-Diffusion (CD): CD updates the most relevant weights related to the input textual features, including W_K and W_V within the cross-attention units, as well as the new token embedding. The implementation details are provided in the supplement.

Evaluation metrics. The synthetic images should faithfully capture the visual characteristics of the input image while accurately conveying all elements of the target text. We employ two key metrics: (1) The image-alignment metric evaluates the reconstruction of concepts, which measures the pairwise CLIP-space cosine similarity [11] between the generated images and the corresponding real images. (2) The text-alignment metric assesses the editing effectiveness of the fine-tuned model by calculating the text-image similarity between the generated images and the provided prompts using CLIP [15]. Notably, these two indicators often conflict with each other [45]. For each concept, we synthesize 16 samples per prompt, using 50 DDIM

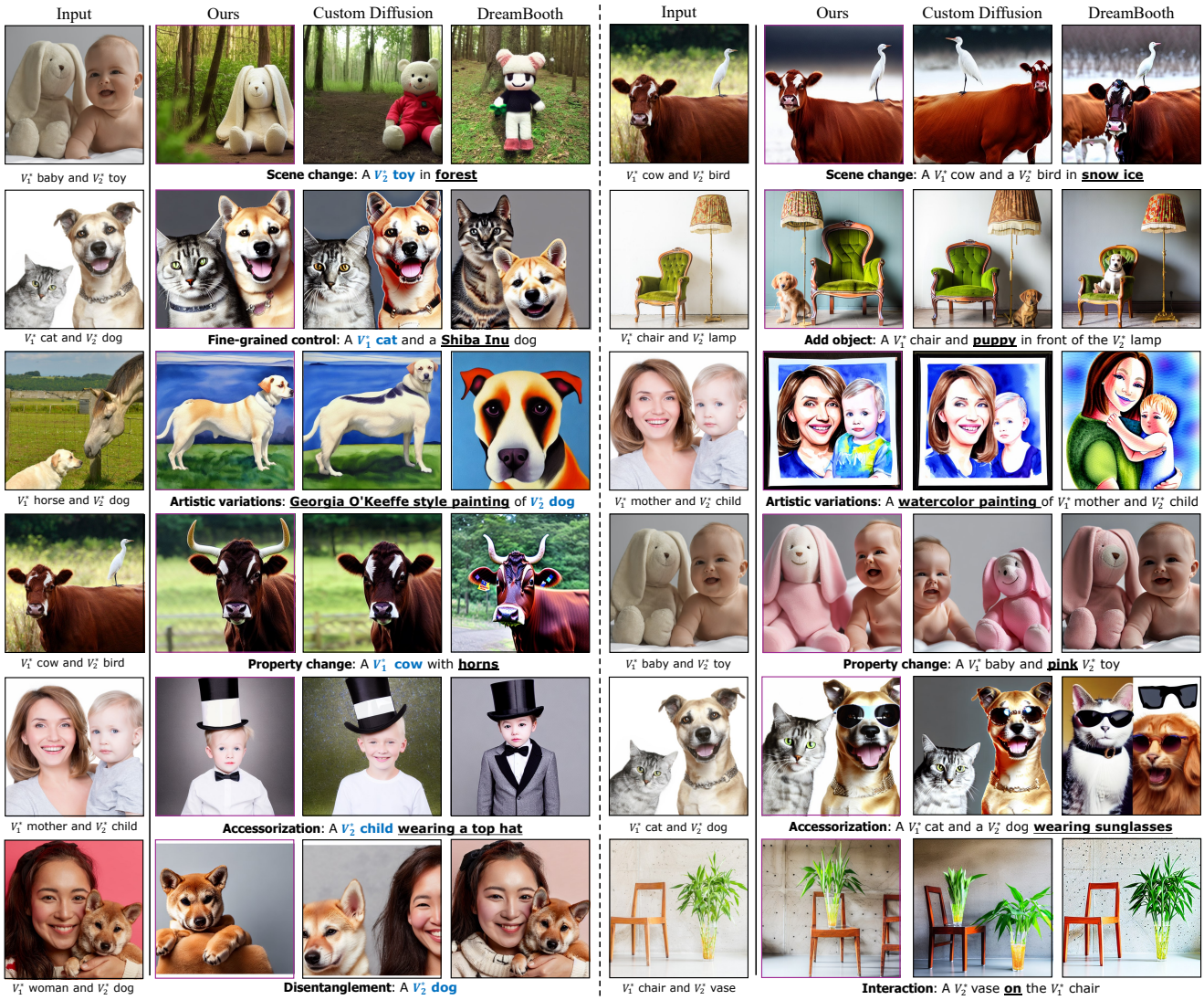

Figure 5. **Qualitative results of independent (left) and combined (right) concepts.** The target prompt in each row represents a distinct context including learned concepts. Our method shows the highest visual similarity to the input image compared to Custom Diffusion and DreamBooth (especially in the first row, the results containing the specific toy) while preserving robust editability. Additionally, we show the ability to address the language drift issue and the disentanglement capability on the left of the second and last row, respectively.

steps and a guidance scale of 6. For comparison, we provide scores for combined concepts, the first concept, the second concept, and their average (referred to as Combined, Concept_1, Concept_2, and Mean in Fig. 6). For instance, if the training image caption is "V_1^* cat and V_2^* dog", the test prompts of the Combined, Concept_1 and Concept_2 settings are "V_1^* cat and V_2^* dog in a garden", "V_1^* cat wearing a hat", "A pink V_2^* dog", respectively. When testing on independent concepts, we calculate the image-alignment metric between the synthesized images and the segmented image containing only the corresponding subject.

Implementation details. We fine-tune the Stable Diffusion [43] model for 250 steps, with a batch size of 8 and a learning rate of 8×10^{-5}. Similar to [23], we employ clip-retrieval [2] to select 200 samples from LAION-5B [42] dataset as regularization images. Captions of these selected images exhibit a similarity of over 0.85 in the CLIP textual embedding space with the input text. Meanwhile, we use the data augmentation in [23]. In our experiments, we apply the proposed cross-attention calibration to the 16×16 attention units, which have been shown to contain the most semantic information [14].

4.2. Comparison Results

Quantitative comparisons. Fig. 6a illustrates the results averaged across ten datasets. As shown, we outperform all the compared methods, especially on the image-alignment scores. Specifically, despite Textual Inversion

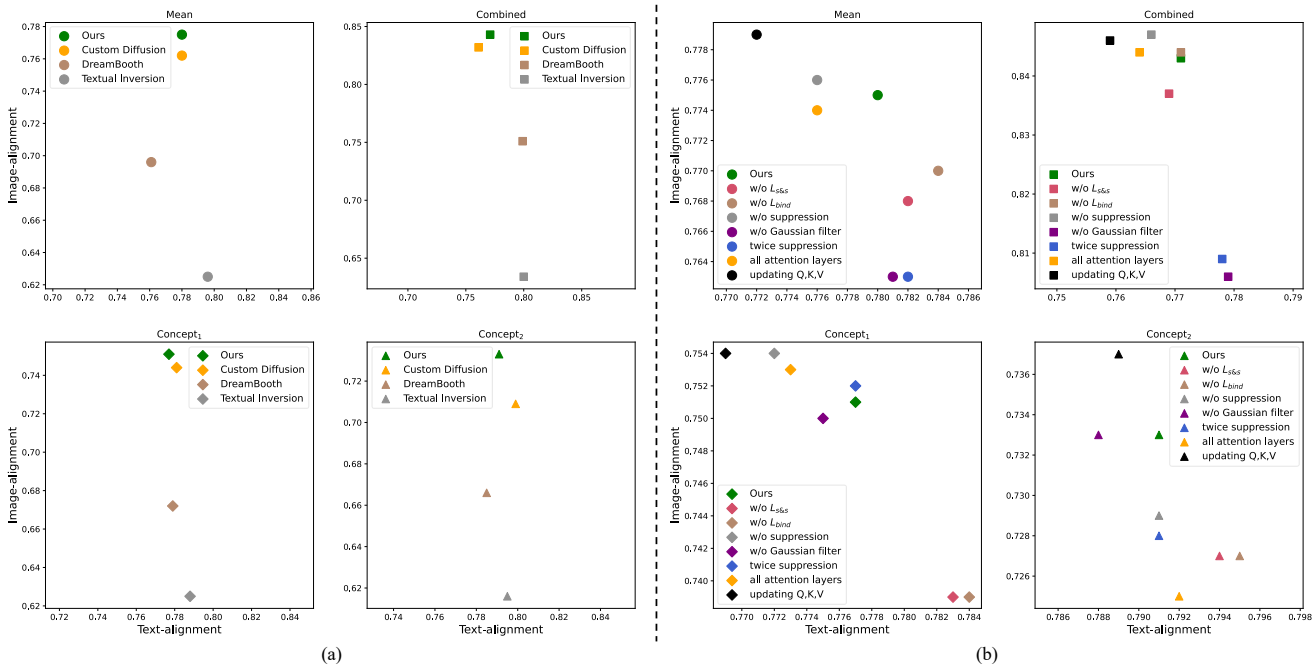

Figure 6. **Quantitative evaluation results.** (a) Compared to state-of-the-art methods, our approach (green) achieves the highest image-alignment score, particularly noticeable in Concept$_2$, while maintaining a text-alignment score similar to that of other methods. (b) Ablation study results. Our full method (green) strikes the best balance between reconstruction and editability.

(TI) achieving the highest text-alignment score, it has the lowest image-alignment score, indicating its struggle to maintain the appearance of concepts. DreamBooth (DB) outperforms TI in image-alignment score but falls significantly short compared to our approach in both metrics. Custom Diffusion (CD) maintains a better balance between the two metrics and competes with ours in combined concepts scores and Concept$_1$ scores. However, there is a noticeable performance gap in the scores for Concept$_2$. In summary, we achieve the highest image fidelity while maintaining strong text editing effectiveness. Detailed results for each dataset can be found in the supplement.

Qualitative comparisons. We visually demonstrate the favorable outcomes in Fig. 5. Concretely, we design diverse target prompts to assess the learned independent concepts and combined concepts in different editing scenarios, including scene changes, object addition, style transfer, property change, accessory addition, interactions between multiple concepts, concept decoupling, and the ability to address the language drift (e.g., generating a specific cat consistent with the input and a dog with a breed distinct from the one present in the input). As shown in Fig. 5, images synthesized by DB either lack key attributes of the concepts or suffer from severe overfitting to the input image. With most of its parameters frozen, CD improves editability and reconstruction compared to DB. However, it still struggles to preserve concepts' appearances or decouple from the input image, especially as shown in the first and last rows of Fig. 5. By incorporating cross-attention calibration, our method achieves high visual fidelity and maintains effective cross-concept disentanglement during T2I generation. For the sake of space efficiency, additional results including Textual Inversion are provided in the supplement.

4.3. Ablation Studies

We conduct ablation studies to show the effectiveness of each component and analyze the influence of different design choices, adopting the same setup described in Sec. 4.1.

To assess the necessity of each component, we set up the following experiment settings: (1) Removing the $\mathcal{L}_{\text{s\&s}}$ loss, (2) removing the $\mathcal{L}_{\text{bind}}$ loss, (3) removing the suppression strategy, (4) removing the Gaussian filter, (5) applying twice suppression (in contrast to one-time). Detailed results are presented in Fig. 6b. As shown, our full model achieves a balanced performance between visual fidelity and editing effectiveness for both combined and independent concepts. Removing either the $\mathcal{L}_{\text{bind}}$ or $\mathcal{L}_{\text{s\&s}}$ loss results in a significant decrease in image-alignment for both Concept$_1$ and Concept$_2$. Similarly, the removal of the Gaussian filter leads to a notable reduction in image-alignment for combined concepts. No suppression significantly harms image-alignment for Concept$_2$, confirming the benefits of sharper boundaries in $A_t^{s_i}$ for understanding multiple concepts (as explained in Sec. 3.3). Meanwhile, this also leads to lower text-alignment for both Concept$_1$ and Concept$_2$. Furthermore, applying twice suppression has detrimental effects on

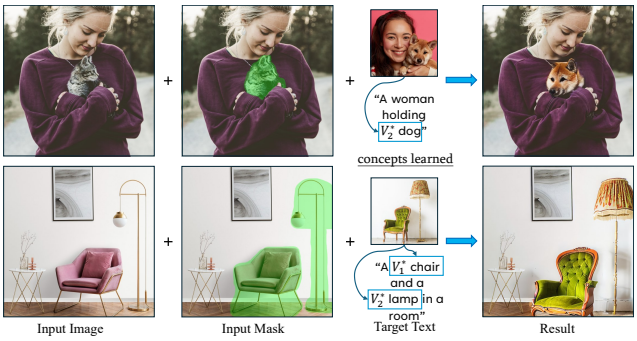

Figure 7. **Applications in image inpainting.** Given an input image and its corresponding mask, our method can seamlessly inpaint the learned concepts into the masked region.

Figure 8. **Integrating with LoRA [18].** Our method can incorporate the LoRA parameters to fully convey the semantics (e.g., enhancing texture details).

image-alignment as it filters out important information.

On the other hand, there are two design choices worth considering. As indicated in [44], averaging all scales of attention layers, instead of just using the 16×16 scale, could potentially yield improved attribution maps for each input word. Therefore, we explore (1) impose constraints on the average of all scales attention layers. Additionally, we investigate releasing more parameters, specifically (2) updating the W_Q, W_K, and W_V matrices within the cross-attention units (in contrast to our approach, which only updates the W_K and W_V). As depicted in Fig. 6b, operating on all scales of attention layers resulted in the model's inability to reconstruct Concept$_2$. Updating W_Q, W_K, and W_V does help the model remember the appearances of concepts but leads to a significant decrease in text-alignment. This suggests that updating more parameters does not preserve the good features of the pre-trained model.

4.4. Applications

Personalized concept inpainting. With any image and its corresponding mask, our method can seamlessly integrate learned concepts into the masked region while preserving the rest of the image, as shown in Fig. 7. Users can effortlessly perform inpainting by simply modifying the text prompt, thanks to our method's conversion of concepts into new word embeddings.

Compatible with LoRA [18]. LoRA techniques, actively discussed in the community, such as CivitAI [6], have gained popularity for enhancing specific capabilities of T2I models, such as improving the ability to refine images. LoRA adds small, trainable parameters to the frozen T2I models for fine-tuning, and our method is orthogonal with it. Therefore, we combine the LoRA with our trained model to unlock a wider range of applications, as shown in Fig. 8. This combination is akin to domain-specific pre-training on a large dataset before personalization [12, 25], with the added benefit of having access to a wealth of readily available LoRA parameters in the community.

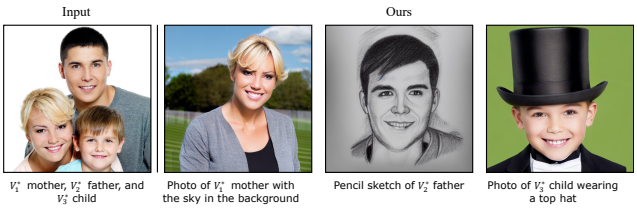

Figure 9. **Applications in extending three concepts.** Enabling edits on three concepts within a single image.

Extending to three concepts. We explore the application of our method to the more challenging task of capturing three concepts from a single image, as shown in Fig. 9. In this scenario, we employ the $\mathcal{L}_{s\&s}$ loss for each pair of the three class tokens to disentangle these concepts.

5. Conclusions and Limitations

We propose the *DisenDiff* to mimic multiple concepts from a single image. We introduce constraints on the cross-attention units to attain precise attention maps for crucial tokens, mitigating the overfitting to the single image and accurately capturing concept appearances. Consequently, our method enables diverse edits involving combined or independent concepts while enhancing the visual similarity between the synthesized images and the input image. Furthermore, we show the flexibility of our method by evaluating several applications.

Limitations. Disentangling fine-grained categories becomes notably challenging when two subjects from the same category co-exist in a single image, such as Golden Retriever and Border Collie dogs. Additionally, while our method can handle images with three concepts, its performance degrades considerably. This can be attributed to the limitations of existing T2I models in such scenarios, as well as the need for algorithm adjustments to address these specific challenges. We believe that there is considerable room to enhance the performance in these complex tasks.

Acknowledgment. This work is supported by Shanghai Science and Technology Program "Federated based cross-domain and cross-task incremental learning" under Grant No. 21511100800, Natural Science Foundation of China under Grant No. 62076094 and No. 62201341.

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 1, 2

[2] Romain Beaumont. Clip retrieval: Easily compute clip embeddings and build a clip retrieval system with them. https://github.com/rom1504/clip-retrieval, 2022. 6

[3] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 2

[4] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 2, 3

[5] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373*, 2023. 3

[6] Civitai. Civitai. https://civitai.com/, 2022. 8

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[8] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022. 2

[9] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *The Eleventh International Conference on Learning Representations*, 2022. 3

[10] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022. 2

[11] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 3, 5

[12] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4):1–13, 2023. 2, 3, 8

[13] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 2

[14] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2022. 3, 4, 6

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021. 5

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[17] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022. 2

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 8

[19] Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*, 2023. 2, 3

[20] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[21] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 2

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[23] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 2, 3, 4, 6

[24] Jason Lee, Kyunghyun Cho, and Douwe Kiela. Countering language drift via visual grounding. *arXiv preprint arXiv:1909.04499*, 2019. 4

[25] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023. 2, 3, 8

[26] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 3

[27] Zhiheng Li, Martin Renqiang Min, Kai Li, and Chenliang Xu. Stylet2i: Toward compositional and high-fidelity text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18197–18207, 2022. 2

[28] Yuchen Lu, Soumye Singhal, Florian Strub, Aaron Courville, and Olivier Pietquin. Countering language drift with seeded iterated learning. In *International Conference on Machine Learning*, pages 6437–6447. PMLR, 2020. 4

[29] Elman Mansimov, Emilio Parisotto, Lei Jimmy Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. 2

[30] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 3

[31] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pages 16784–16804. PMLR, 2022. 2

[32] Dong Huk Park, Grace Luo, Clayton Toste, Samaneh Azadi, Xihui Liu, Maka Karalashvili, Anna Rohrbach, and Trevor Darrell. Shape-guided diffusion with inside-outside attention. *arXiv preprint arXiv:2212.00210*, 2022. 3

[33] Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3

[35] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 2

[36] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

[37] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1, 2

[38] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016. 2

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3

[40] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1, 3, 4

[41] 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. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 2

[42] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2, 6

[43] Stable diffusion. https://huggingface.co/runwayml/stable-diffusion-v1-5, 2022. 3, 6

[44] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the daam: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885*, 2022. 8

[45] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2, 3, 5

[46] Dani Valevski, Matan Kalman, Eyal Molad, Eyal Segalis, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning a diffusion model on a single image. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 3

[47] Ruichen Wang, Zekang Chen, Chen Chen, Jian Ma, Haonan Lu, and Xiaodong Lin. Compositional text-to-image synthesis with attention map control of diffusion models. *arXiv preprint arXiv:2305.13921*, 2023. 3

[48] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022. 3

[49] Zhe Wang, Qida Dong, Wei Guo, Dongdong Li, Jing Zhang, and Wenli Du. Geometric imbalanced deep learning with feature scaling and boundary sample mining. *Pattern Recognition*, 126:108564, 2022. 2

[50] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 2, 3

[51] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 516–520, 2016. 4

[52] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 2

[53] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842, 2021. 2

[54] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3

[55] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6027–6037, 2023. 3

[56] Xiaojin Zhu, Andrew B Goldberg, Mohamed Eldawy, Charles R Dyer, and Bradley Strock. A text-to-picture synthesis system for augmenting communication. In *AAAI*, pages 1590–1595, 2007. 2