Textual Localization: Decomposing Multi-concept Images for Subject-Driven Text-to-Image Generation

Junjie Shentu [1] **Matthew Watson** [1] **Noura Al Moubayed** [1]

Abstract

Subject-driven text-to-image diffusion models empower users to tailor the model to new concepts absent in the pre-training dataset using a few sample images. However, prevalent subject-driven models primarily rely on single-concept input images, facing challenges in specifying the target concept when dealing with multi-concept input images. To this end, we introduce a textual localized text-to-image model (*Texual Localization*) to handle multi-concept input images. During fine-tuning, our method incorporates a novel cross-attention guidance to decompose multiple concepts, establishing distinct connections between the visual representation of the target concept and the identifier token in the text prompt. Experimental results reveal that our method outperforms or performs comparably to the baseline models in terms of image fidelity and image-text alignment on multi-concept input images. In comparison to *Custom Diffusion*, our method with hard guidance achieves CLIP-I scores that are 7.04%, 8.13% higher and CLIP-T scores that are 2.22%, 5.85% higher in single-concept and multi-concept generation, respectively. Notably, our method generates cross-attention maps consistent with the target concept in the generated images, a capability absent in existing models.

1. Introduction

Diffusion models have demonstrated unprecedented capabilities in generating high-quality and diverse images, effectively addressing the mode collapse problem encountered by Generative Adversarial Networks (GANs) (Ho et al., 2020; Dhariwal & Nichol, 2021). By leveraging cross-attention layers within the UNet architecture (Ronneberger et al.,

[1]Department of Computer Science, Durham University, Durham, UK. Correspondence to: Noura Al Moubayed <noura.al-moubayed@durham.ac.uk>.

Data and code are released at https://github.com/junjie-shentu/Textual-Localization.

2015), diffusion models can be conditioned on information in different modalities. Text-to-image diffusion models, a specific subclass of conditional diffusion models, exemplify remarkable proficiency in producing images semantically aligned with natural language prompts (Nichol et al., 2021; Rombach et al., 2022; Ramesh et al., 2022; Gu et al., 2022).

Despite the strong image-text connections established by text-to-image models, introducing new concepts not present in the pre-training datasets remains challenging (Gal et al., 2022). In response, studies aimed at "customizing" text-to-image models for generalization to newly introduced concepts propose fine-tuning pre-trained models with a few (typically 3 to 5) images of the target object, defining them as subject-driven text-to-image models (Gal et al., 2022; Ruiz et al., 2023a; Kumari et al., 2023; Li et al., 2023; Jia et al., 2023; Gal et al., 2023; Arar et al., 2023; Ruiz et al., 2023b; Ma et al., 2023a). However, prevalent subject-driven models are designed to learn from images containing a single new concept (termed single-concept images), imposing high requirements for data preparation and demanding a prolonged fine-tuning process when introducing multiple concepts (Kumari et al., 2023). Conversely, the feasibility of applying images containing multiple concepts (termed multi-concept images) for fine-tuning has not been thoroughly explored.

To this end, we evaluate the state-of-the-art model *Custom Diffusion* (Kumari et al., 2023) on multi-concept images. The test involves utilizing the text prompt "*Photo of a $\mathcal{V}$ [class]*", where $\mathcal{V}$ serves as a unique identifier token representing the target concept, and [class] denotes the class name of the concept. We find that the model tends to generate all concepts present in the input images, disregarding the specified target concept in the text prompt. Example instances of these failure cases are showcased in Figure 1.

To address the identified deficiencies in existing subject-driven models, we propose a novel approach named *Textual Localization*. This method aims to decompose input images and achieve precise customization of the target concepts, especially when confronted with multi-concept images. During model fine-tuning, we introduce a cross-attention guidance mechanism that incorporates a new cross-attention loss L_{attn}, which is designed to pinpoint the target concept re-

gion in the input image and establish a distinct connection between the visual representation of the target concept and the identifier token $\mathcal{V}$. Our method encompasses two distinct strategies of cross-attention guidance: hard guidance and soft guidance, both of which eliminate the model's attention on non-target concepts but apply different ways of attention activation in the target region of the input images. The cross-attention guidance manipulates the cross-attention maps between the input images and identifier token $\mathcal{V}$, thereby influencing the model's attention. Additionally, we compile a dataset comprising both multi-concept images and single-concept images for model training and evaluation. Experimental results indicate that our method either outperforms or matches baseline models in both single-concept and multi-concept generation when taking multi-concept images as input. Moreover, our approach explicitly showcases the connection between the visual representation of the target concept and identifier token $\mathcal{V}$ through the cross-attention maps.

Figure 1. Failure cases in single-concept generation by Custom Diffusion when fine-tuning on multi-concept inputs

2. Related Work

2.1. Text-to-image diffusion model

Diffusion model, functioning as a likelihood-based model, attains state-of-the-art performance in generating high-quality images, surpassing other generative models (Ho et al., 2020; Dhariwal & Nichol, 2021). Additionally, the inclusion of cross-attention layers equips the diffusion model with the capability to incorporate conditioning information in diverse modalities (Rombach et al., 2022), with natural language being one of the predominant sources of conditioning information. Nichol et al. (2021) and Saharia et al. (2022) apply a text-to-image generation model in the pixel space with classifier-free guidance (Ho & Salimans, 2022). They utilize a transformer (Vaswani et al., 2017) and a pre-trained large language model (Ramesh et al., 2021) as text encoders, respectively. Besides, Rombach et al. (2022) train a diffusion model in the latent space by using a Variational Autoencoder (VAE) (Kingma & Welling,

2013) to project images into latent space, using a pre-trained text encoder from Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) to process the text prompt. Ramesh et al. (2022) propose a multimodal latent space by training a prior model to generate the CLIP image embedding of the input text prompt, and generating images conditioned on the image embedding.

2.2. Subject-driven text-to-image generation

To address the challenge of generating novel concepts absent in the pre-training dataset, subject-driven generation is devised to customize the text-to-image generation model using a limited set of sample images (Ruiz et al., 2023a). In subject-driven generation, rare-occurring identifier tokens from the vocabulary are inserted into the text prompt to establish a connection with the input images during the fine-tuning process. Various training targets are explored, including solely the text encoder (Gal et al., 2022), the text encoder along with cross-attention layers in the UNet (Kumari et al., 2023; Shi et al., 2023), and the text encoder together with the entire UNet (Ruiz et al., 2023a). Additionally, Li et al. (2023) and Jia et al. (2023) incorporate an image encoder to obtain more accurate and robust embeddings of the input images, replacing the identifier tokens. In a bid to further optimize the fine-tuning process and reduce the number of training parameters, Arar et al. (2023) and Ruiz et al. (2023b) propose applying Low-Rank adaptation (LoRA) (Hu et al., 2021) for expedited personalization. However, the majority of the aforementioned models predominantly focus on learning from single-concept images, while our approach excels in decomposing input images and facilitating precise learning from multi-concept images.

2.3. Cross-attention in text-to-image image generation

Diffusion models harness the cross-attention layers (Vaswani et al., 2017) embedded in the underlying UNet backbone to integrate conditioning information from text prompts into the generated images (Rombach et al., 2022). These cross-attention layers amalgamate information from both images and text, producing cross-attention maps that represent the probability distribution over text tokens for each image patch in the latent space (Tang et al., 2022; Chefer et al., 2023). Guiding these cross-attention maps during inference empowers the pre-trained diffusion model to generate images with superior semantic alignment to the provided text prompts (Feng et al., 2022; Chefer et al., 2023; Wang et al., 2023; Phung et al., 2023), achieve image editing (Hertz et al., 2022), and provide positional control over the contents in the generated images (Ma et al., 2023b; He et al., 2023; Chen et al., 2023a; Phung et al., 2023). Besides, cross-attention guidance is also applied during training to help achieve zero-shot personalized image generation, although the generation quality is inferior compared to the models

fine-tuned on input images (Ma et al., 2023a). Xiao et al. (2023) utilize cross-attention guidance to address the identity blending problem and enable multi-subject generation. Notably, their model's performance is demonstrated on a human face dataset, with its general subject performance remaining undisclosed.

3. Textual Localized Diffusion Model

3.1. Preliminaries

In this study, we adopt Stable Diffusion (SD) as the foundational model, built upon the Latent Diffusion Model (LDM) (Rombach et al., 2022). For an input image $x \in \mathbb{R}^{H \times W \times 3}$, SD initially projects x into a latent representation $z \in \mathbb{R}^{h \times w \times c}$ by employing the encoder $\mathcal{E}$ of a VAE (Kingma & Welling, 2013), where c denotes the latent feature dimension. The downsampling follows a factor $f = H/h = W/w$, determining the downsampling scale. The diffusion process is subsequently executed on the latent representation by introducing noise into z, forming a fixed-length Markov Chain denoted as $z_1 \ldots z_T$, where T signifies the length of the chain. SD trains the UNet to learn the reverse process of the Markov Chain, predicting a denoised variant of the input z_t given the timestep $t \in [1, T]$. In the context of text-to-image generation, the conditioning information y from the text prompts is projected into an intermediate representation $\tau_\theta(y)$, where τ_θ is a pre-trained CLIP text encoder. The training objective of the text-to-image diffusion model can be expressed as:

$$L_{LDM} = \mathbb{E}_{\mathcal{E}(x),y,t} \left[\|\varepsilon - \varepsilon_\theta \left(z_t, t, \tau_\theta(y) \right)\|_2^2 \right] \quad (1)$$

where ε and ε_θ represent the standard Gaussian noise ($\varepsilon \sim \mathcal{N}(0, 1)$) and predicted noise residue, respectively. Specifically, the intermediate representation $\tau_\theta(y)$ is linked to the intermediate layers of the UNet through cross-attention layers using the following mapping:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}} \cdot V \right) \quad (2)$$

$$Q = W_Q^{(i)} \cdot \varphi_i(z_t), K = W_K^{(i)} \cdot \tau_\theta(y), V = W_V^{(i)} \cdot \tau_\theta(y) \quad (3)$$

where d signifies the output dimension of the query (Q) and key (K) features. $W_Q^{(i)}$, $W_K^{(i)}$, and $W_V^{(i)}$ are learnable projection matrices in cross-attention layer i, $\varphi_i(z_t)$ is a flattened intermediate representation of the noisy latent z_t. The cross-attention map at layer i is given by:

$$Attn^{(i)} = softmax\left(\frac{Q^{(i)} K^{(i)T}}{\sqrt{d}} \right) \quad (4)$$

3.2. Pipeline of Textual Localization

Subject-driven text-to-image models establish a connection between the new concept from the input images and the identifier token $\mathcal{V}$. During the fine-tuning process, the text embedding of $\mathcal{V}$ is refined to represent the target concept through the cross-attention layers, and the model learns a connection between the text embedding of $\mathcal{V}$ with the visual representation of the target concept in pixel space (Ruiz et al., 2023a; Kumari et al., 2023). However, when presented with multi-concept images, this connection becomes ambiguous, as depicted in Figure 1. To address this ambiguity, we enhance the model by incorporating additional cross-attention guidance during the fine-tuning process. Our proposed method is denoted as the textual localized text-to-image model, or *Textual Localization*.

A single fine-tuning step of *Textual Localization* is illustrated in Figure 2. Following *Custom Diffusion*, we only optimize the text encoder as well as the W_K and W_V matrices in the cross-attention layers in the UNet. The learning objective of the *Textual Localization* involves minimizing a loss function comprised of a denoising loss $L_{denoise}$ and an attention loss L_{attn}. The denoising loss encompasses the original training objective of LDM, as given by Equation (1), along with a class-specific prior preservation loss, expressed as:

$$L_{prior} = \mathbb{E}_{\mathcal{E}(x_{pr}),y_{pr},t} \left[\|\varepsilon - \varepsilon_\theta \left(z_{pr_t}, t, \tau_\theta(y_{pr}) \right)\|_2^2 \right] \quad (5)$$

where x_{pr} is the sample generated by the pre-trained text-to-image model under the text prompt y_{pr} that solely specifies the class name of the target concept. The inclusion of the class-specific prior preservation loss serves to maintain output diversity and prevent language drift (Ruiz et al., 2023a). The cross-attention loss L_{attn} is formulated to bias the model's attention, establishing a clear connection between the identifier token $\mathcal{V}$ and the target concept. The intricacies of cross-attention guidance are elucidated in Section 3.3. Consequently, the overarching training objective is to minimize:

$$L = L_{LDM} + \lambda L_{prior} + \delta L_{attn} \quad (6)$$

where λ and δ are two scaling coefficients.

3.3. Cross-attention Guidance

The positional information of the target concept in the input images is derived from the segmentation maps, generated by a pre-trained segmentation model. Through the utilization of these segmentation maps, we introduce two distinct strategies for cross-attention guidance: hard guidance and soft guidance.

Hard guidance. In the case of hard guidance, the cross-attention map of the identifier token $Attn_{\mathcal{V}}$ is optimized to align with the segmentation map $Seg \in \mathbb{R}^{H' \times W'}$. The attention loss L_{attn} is computed as the mean square error (MSE) between $Attn_{\mathcal{V}}$ and Seg. Thus, L_{attn} can be formu-

Figure 2. Illustration of a single step of the fine-tuning process of *Textual Localization*

lated as:

$$L_{attn} = \frac{1}{H'W'}\sum_{i=1}^{H'}\sum_{j=1}^{W'}\left(Seg(i,j) - Attn_{\mathcal{V}}(i,j)\right)^2 \tag{7}$$

Note that L_{attn} is calculated in the pixel space. Given that the cross-attention maps $Attn^{(i)}$ extracted from various layers possess distinct resolutions determined by their positions in the UNet, all cross-attention maps of the identifier token $Attn_{\mathcal{V}}^{(i)}$ are up-scaled to $H' \times W'$ and subsequently averaged to obtain $Attn_{\mathcal{V}}$.

Soft guidance. The objective of soft guidance is to eliminate the model's attention on regions outside of the target concept in the input images, without influencing attention within the target concept region. The attention loss L_{attn} is formulated as the element-wise product of a binary matrix (B_{Inv}), representing the inverse segmentation map, and the MSE between $Attn_{\mathcal{V}}$ and Seg, which is given by:

$$L_{attn} = \frac{1}{H'W'}\sum_{i=1}^{H'}\sum_{j=1}^{W'}$$
$$\left[(Seg(i,j) - Attn_{\mathcal{V}}(i,j))^2 \cdot B_{Inv}(i,j)\right] \tag{8}$$

$$B_{Inv}(i,j) = \begin{cases} 1 & \text{if } Seg(i,j) = 0 \\ 0 & \text{if } Seg(i,j) > 0 \end{cases} \tag{9}$$

While both hard guidance and soft guidance effectively reduce the model's attention on non-target concepts, they diverge in their treatment of the target concept region in the input images. Notably, hard guidance influences the model to activate attention in the target region, producing cross-attention maps that align with the segmentation map. In contrast, soft guidance does not alter attention activation in the region of the target concept.

4. Experiments and Results

4.1. Experimental setup

Datasets. We curated a dataset comprising 10 novel concepts encompassing general and everyday items. To assess the model's efficacy with multi-concept images, we randomly formed five groups by pairing two concepts. Additionally, we prepared single-concept images for each concept to facilitate evaluation. Each single and multiple concept set consists of five images, and samples can be found in Appendix A. To pinpoint the locations of the target concepts in the input images, we employed the Grounding DINO detection model (Liu et al., 2023) to generate bounding boxes. These bounding boxes served as cues for the segmentation model SAM (Kirillov et al., 2023) to derive segmentation maps for the target concepts.

Baseline models. We conduct a comparative analysis of our method against two baseline models, namely, *DreamBooth* (Ruiz et al., 2023a) and *Custom Diffusion* (Kumari et al., 2023). Both baseline models are subject-driven text-to-image models and integrate the class-specific prior preservation loss to mitigate language drift. *DreamBooth* undertakes fine-tuning of the text encoder and the entire UNet, while *Custom Diffusion* focuses solely on optimizing the text encoder and the W_K and W_V matrices within the cross-attention layers of the UNet.

Evaluation metrics. We evaluate our method on the benchmark raised by the baseline models to present a fair comparison, focusing on image fidelity and image-text alignment. To gauge image fidelity, we compute the cosine similarity between the CLIP embeddings of the generated images and the real images, denoted as CLIP-I. Additionally, we calculate the Kernel Inception Distance (KID) (Bińkowski et al., 2018) between the generated and real images, providing further evidence of image fidelity. Furthermore, the average Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) is computed for the generated images of the same target concept under identical text prompts. This measure reveals the diversity of the generated images and aids in assessing potential model overfitting. For image-text alignment, the cosine similarity between CLIP embeddings of the generated images and the corresponding text prompts is calculated (CLIP-T). Note that when evaluating CLIP-T, the identifier token $\mathcal{V}$ is omitted, as the CLIP text encoder has not undergone fine-tuning on $\mathcal{V}$.

Implementation details. For all experiments, we employ SD v1.5 trained on the LAION-5B dataset (Schuhmann et al., 2022) as the foundational model. Prior to fine-tuning, we leverage the pre-trained SD to generate 200 images per target concept for computing L_{prior}, using text prompts consisting of the class name of the respective target concept. In the case of our method, the scaling coefficients λ

Figure 3. Qualitative comparison in single-concept generation

and δ are set to 1.0. We extract cross-attention maps from cross-attention layers with downsampling factors $f \in 2, 4, 8$ (Hertz et al., 2022; Chefer et al., 2023), and up-scale to $H' \times W'$, where $H' = W' = 256$ for computational efficiency. The identifier token $\mathcal{V}$ ($\mathcal{V}_1$ and $\mathcal{V}_2$ for multiple concepts) is initialized with the token ID 48136 in the pretrained CLIP tokenizer (Gal et al., 2022; Ruiz et al., 2023a). The learning rate is set to 1.0×10^{-5} for our method and *Custom Diffusion*, and 5.0×10^{-6} for *DreamBooth*. Maintaining a fixed batch size of 2, all models are fine-tuned on an NVIDIA A100 GPU.

4.2. Single-concept generation

We assess each model's capability to generate a single new concept when provided with multi-concept images as input. For each target concept, we employ 10 text prompts and generate 50 image samples per prompt, resulting in a total of 500 images. Additionally, we evaluate the performance of baseline models when taking single-concept images as input for a more comprehensive analysis. The average model performance across all target concepts is summarized in Table 1. Our method achieves superior performance in CLIP-T and LPIPS, with the soft guidance

variant obtaining the highest score and the hard guidance variant securing the second-best score. This indicates that our method effectively preserves more semantic information from the text prompts and exhibits reduced overfitting. On the other hand, *DreamBooth*, fine-tuned on single-concept images, attains the highest scores in CLIP-I and KID, reflecting higher similarities between input and generated images. However, it is noteworthy that these results may be biased, as baseline models fine-tuned on single-concept images also use the same images for evaluation. Moreover, *DreamBooth*, optimizing more parameters, showcases enhanced learning of visual representation but also accelerates the loss of learned prior knowledge, resulting in lower CLIP-T and LPIPS scores. Furthermore, compared to *Custom Diffusion*, our method exhibits an overall improvement, as the CLIP-I, CLIP-T, and LPIPS scores of the hard guidance variant are 7.04%, 2.22%, 0.91% higher, and the KID score is 6.10% lower on multi-concept images.

The qualitative evaluation results are presented in Figure 3. Our method, fine-tuned on multi-concept images, successfully generates images containing only the target concept while retaining rich semantic information from the text prompt. In contrast, both baseline models encounter dif-

Figure 4. Qualitative comparison in multi-concept generation

ficulties in clarifying the target concept and are prone to generating images containing all concepts present in the input images. Additionally, while *DreamBooth* demonstrates a better visual representation of the input concept, it tends to lose more semantic knowledge, leading to results inconsistent with the text prompts (e.g., $\mathcal{V}$ *doll under water*). Further examples and analysis can be found in Appendix B.1.

4.3. Multi-concept generation

We evaluate each model's performance in generating multiple new concepts when fine-tuning them on multi-concept images. Two identifier tokens, $\mathcal{V}_1$ and $\mathcal{V}_2$, are introduced in the text prompts to represent the two target concepts. All target concepts are learned simultaneously during fine-tuning, pairing the input images with the text prompt "*photo of a $\mathcal{V}_1$ [class$_1$] and a $\mathcal{V}_2$ [class$_2$]*". Baseline models are also jointly fine-tuned on single-concept images of the two target concepts, following the approach by Kumari et al. (2023). We generate 50 images per prompt for 10 text prompts in each multi-concept group, resulting in a total of 500 images. Quantitative evaluation results are presented in Table 1.

Overall, our method exhibits superior performance in multi-concept generation. The hard guidance variant achieves the best CLIP-I and KID scores as well we the second-best CLIP-T score, while the soft guidance variant records the highest LPIPS score. In comparison to *Custom Diffusion*, the CLIP-I, CLIP-T, and LPIPS scores of the hard guidance variant are 8.13%, 5.85%, 7.50% higher, and the KID score is 7.75% lower on multi-concept images. It's important to note that as both fine-tuning and evaluation use multi-concept images, there might be bias in the results when comparing our method with baseline models fine-tuned on single-concept images. Nevertheless, our method outperforms both baseline models fine-tuned on multi-concept images in terms of CLIP-I, KID, and LPIPS. Notably, while *Custom Diffusion* fine-tuned on single-concept images attains the highest CLIP-T score, it exhibits the worst CLIP-I and KID scores, indicating insufficient learning of visual representation from input images, resulting in less loss of semantic knowledge during fine-tuning.

Results of the qualitative evaluation in multi-concept generation are presented in Figure 4. All models demonstrate the

Table 1. Comparison between our method and baseline models (Bold indicates the best value, underline represents the second-best value)

Method	Input Concept	Single-concept generation				Multi-concept generation			
		CLIP-I ↑	CLIP-T ↑	KID ↓	LPIPS ↑	CLIP-I ↑	CLIP-T ↑	KID ↓	LPIPS ↑
DreamBooth	Single	**0.6583**	0.2161	**0.1297**	0.5884	<u>0.5189</u>	0.1970	0.1952	0.5971
Custom Diffusion	Single	0.5525	0.2645	<u>0.1865</u>	0.6355	0.4600	**0.2890**	0.2179	<u>0.6485</u>
DreamBooth	Multiple	<u>0.5999</u>	0.2099	0.1985	0.6162	0.5034	0.2136	<u>0.1625</u>	0.6036
Custom Diffusion	Multiple	0.4883	0.2612	0.2228	0.6595	0.4907	0.2548	0.1743	0.5890
Ours (hard guidance)	Multiple	0.5227	<u>0.2670</u>	0.2092	<u>0.6655</u>	**0.5306**	<u>0.2697</u>	**0.1608**	0.6332
Ours (soft guidance)	Multiple	0.5077	**0.2680**	0.2205	**0.6685**	0.4951	0.2638	0.1781	**0.6508**

capability to generate both target concepts when fine-tuned on multi-concept images. However, *DreamBooth* exhibits a more pronounced loss of semantic knowledge, leading to weaker representations of text prompts in some sample images. For instance, with the text prompt "$\mathcal{V}_1$ *helmet and* $\mathcal{V}_2$ *headphone at a beach with a view of the seashore*", only a small amount of water is presented. Similarly, a water-color painting is shown on the penbag with the text prompt "*A watercolor painting of* $\mathcal{V}_1$ *pot and* $\mathcal{V}_2$ *penbag*". Furthermore, baseline models trained on single-concept images display property fusion of the two target concepts or only depict one concept in generated images. In contrast, our method appropriately presents both target concepts while accurately reflecting semantic information from the text prompts. Additional examples and detailed results are provided in Appendix B.2.

4.4. Probing into cross-attention maps

A crucial aspect of subject-driven models is the establishment of a connection between the identifier token $\mathcal{V}$ and the visual representation of the target concept. However, this connection heavily relies on the diffusion models' perceptual ability to accurately locate the target region. When confronted with multi-concept images, the connection becomes ambiguous due to a lack of guidance, resulting in the presence of non-target concepts in the generated images. To showcase the connection, we extract cross-attention maps from the 16×16 ($f = 4$) layers of the UNet throughout all timesteps during inference, and then upscale to 256×256, which are displayed in Figure 5.

It can be observed that both *DreamBooth* and *Custom Diffusion* tend to produce the non-target concept in single-concept generation, and the cross-attention maps of the identifier token outline the shape of all concepts present in the generated images in both single and multiple concept generation. Moreover, a comparative analysis between the soft guidance variant and the hard guidance variant of our method shows that while both variants avoid generating the non-target concept, the former only partially depicts the outline of the target concept in the cross-attention maps in single-concept generation. This partial connection ex-

Figure 5. Images samples and cross-attention maps of identifier tokens generated by adopted models fine-tuned on multi-concept input images

plains the higher CLIP-T score but lower CLIP-I score of the soft guidance variant compared to the hard guidance variant. In multi-concept generation, the soft guidance variant struggles to differentiate cross-attention maps for different concepts, whereas the hard guidance variant accurately depicts the outlines of each target concept. Consequently, ambiguities persist when using the soft guidance variant for multi-concept generation, leading to lower CLIP-I and CLIP-T scores compared to the hard guidance variant.

4.5. Ablation study

As detailed in Table 1, *DreamBooth* achieves a higher CLIP-I score but a lower CLIP-T score compared to *Custom Diffusion* when provided with the same input images as *DreamBooth* optimizes more model parameters, resulting in a more pronounced loss of prior semantic knowledge. Determining an optimal set of trainable parameters is crucial for

improving the model's ability to learn visual representation while retaining semantic knowledge. To explore this, we conducted an ablation study on parameter optimization.

Building upon the conclusion by Kumari et al. (2023), we delve deeper by evaluating the rate of weight change for the W_Q, W_K, and W_V matrices in the cross-attention layers. We fine-tuned the cross-attention layers on six single concepts from our dataset, calculating the mean rate of weight change for each layer using $\Delta_l = ||\theta_l' - \theta_l||/||\theta_l||$, where θ_l and θ_l' represent weights of the parameters in layer l before and after fine-tuning (Li et al., 2020). The rates of weight change for different matrices with fine-tuning steps are presented in Figure 6. Notably, the weights of W_V undergo the most significant change, while the weights of W_Q and W_K exhibit less pronounced changes. Given this observation, we select three sets of model parameters for optimization: (1) $W_Q + W_K + W_V$, (2) $W_Q + W_V$, and (3) $W_K + W_V$ (adopted in our method). We fine-tune these three parameter sets within the framework of our method with hard guidance and evaluate the generated images. Table 2 showcases the results of the ablation study, revealing that optimizing $W_K + W_V$ achieves the best or second-best performance across most metrics in single-concept generation. Remarkably, it attains both the highest CLIP-I and CLIP-T scores in multi-concept generation. Therefore, optimizing $W_K + W_V$ emerges as a rational choice in this study. Additional details are provided in Appendix B.3.

Figure 6. Rates of weight change of different matrices in cross-attention layers

Figure 7. Failure cases on subject-driven text-to-image generation. Left: failure to capture the details of the cat figurine; Right: failure to generate all concepts in the text prompt. (Text prompts: $\mathcal{V}_1$ *cat figurine* (left) / $\mathcal{V}_1$ *pot and* $\mathcal{V}_2$ *penbag* (right) *in woods with falling leaves in the background*)

Table 2. Ablation study of different selections of trainable parameters (Bold indicates the best value, underline represents the second-best value)

Parameter Set	CLIP-I ↑	CLIP-T ↑	KID ↓	LPIPS ↑
Single-concept generation				
$W_Q + W_K + W_V$	**0.5288**	0.2629	**0.1985**	0.6587
$W_Q + W_V$	0.5142	<u>0.2658</u>	0.2103	<u>0.6601</u>
$W_K + W_V$	<u>0.5227</u>	**0.2670**	<u>0.2092</u>	**0.6655**
Multi-concept generation				
$W_Q + W_K + W_V$	<u>0.5224</u>	0.2644	**0.1512**	<u>0.6313</u>
$W_Q + W_V$	0.5133	<u>0.2659</u>	<u>0.1583</u>	0.6303
$W_K + W_V$	**0.5306**	**0.2697**	0.1608	**0.6332**

5. Discussion and Conclusion

This study introduces a novel subject-driven text-to-image model, termed *Textual Localization*, aimed at mitigating ambiguities inherent in subject-driven models on multi-concept input images. The proposed method incorporates a novel cross-attention guidance to disentangle multiple concepts from input images and establish accurate connections between the visual representations of the target concept and the identifier token in the text prompt. Our method demonstrates superior or comparable performance to baseline models in terms of image fidelity and image-text alignment on multi-concept input images, as the hard guidance variant achieves CLIP-I scores that are 7.04%, 8.13% higher, and CLIP-T scores that are 2.22%, 5.85% higher than *Custom Diffusion* in single-concept and multi-concept generation, respectively. Notably, our technique effectively delineates the outlines of target concepts in cross-attention maps.

Nevertheless, limitations emerge in capturing intricate details of target concepts as shown in Figure 7. Additionally, failure cases may arise in multi-concept generation, where only one concept is generated despite models being fine-tuned on multi-concept images, as depicted in Figure 7. Hence, our focus in future work will be on addressing these limitations. Specifically, we propose adopting more powerful feature extractors (Chen et al., 2023b) to accentuate details in the input images, and integrating guiding techniques during inference (Chefer et al., 2023; Chen et al., 2023a) to ensure the successful generation of all target concepts mentioned in the text prompt.

Impact Statements

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Arar, M., Gal, R., Atzmon, Y., Chechik, G., Cohen-Or, D., Shamir, A., and Bermano, A. H. Domain-agnostic tuning-encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06925*, 2023.

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., and Cohen-Or, D. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023.

Chen, M., Laina, I., and Vedaldi, A. Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373*, 2023a.

Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., and Zhao, H. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023b.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Feng, W., He, X., Fu, T.-J., Jampani, V., Akula, A., Narayana, P., Basu, S., Wang, X. E., and Wang, W. Y. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Gal, R., Arar, M., Atzmon, Y., Bermano, A. H., Chechik, G., and Cohen-Or, D. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4):1–13, 2023.

Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., and Guo, B. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10696–10706, 2022.

He, Y., Salakhutdinov, R., and Kolter, J. Z. Localized text-to-image generation for free via cross attention control. *arXiv preprint arXiv:2306.14636*, 2023.

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Jia, X., Zhao, Y., Chan, K. C., Li, Y., Zhang, H., Gong, B., Hou, T., Wang, H., and Su, Y.-C. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*, 2023.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Kumari, N., Zhang, B., Zhang, R., Shechtman, E., and Zhu, J.-Y. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1931–1941, 2023.

Li, D., Li, J., and Hoi, S. C. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023.

Li, Y., Zhang, R., Lu, J., and Shechtman, E. Few-shot image generation with elastic weight consolidation. *arXiv preprint arXiv:2012.02780*, 2020.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

Ma, J., Liang, J., Chen, C., and Lu, H. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. *arXiv preprint arXiv:2307.11410*, 2023a.

Ma, W.-D. K., Lewis, J., Kleijn, W. B., and Leung, T. Directed diffusion: Direct control of object placement through attention guidance. *arXiv preprint arXiv:2302.13153*, 2023b.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Phung, Q., Ge, S., and Huang, J.-B. Grounded text-to-image synthesis with attention refocusing. *arXiv preprint arXiv:2306.05427*, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. 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*, pp. 22500–22510, 2023a.

Ruiz, N., Li, Y., Jampani, V., Wei, W., Hou, T., Pritch, Y., Wadhwa, N., Rubinstein, M., and Aberman, K. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023b.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., 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.

Shi, J., Xiong, W., Lin, Z., and Jung, H. J. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023.

Tang, R., Liu, L., Pandey, A., Jiang, Z., Yang, G., Kumar, K., Stenetorp, P., Lin, J., and Ture, F. What the daam: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885*, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, R., Chen, Z., Chen, C., Ma, J., Lu, H., and Lin, X. Compositional text-to-image synthesis with attention map control of diffusion models. *arXiv preprint arXiv:2305.13921*, 2023.

Xiao, G., Yin, T., Freeman, W. T., Durand, F., and Han, S. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Appendix

A. Self-constructed Dataset

Our experiments encompass evaluations on both single-concept and multi-concept images. Existing datasets for subject-driven text-to-image models, such as the DreamBooth Dataset (Ruiz et al., 2023a) and CustomConcept101 (Kumari et al., 2023), primarily comprise single-concept images. However, the availability of datasets featuring multi-concept images is limited. To address this, we curated a self-constructed dataset containing both single-concept and multi-concept images, each depicting various general and everyday items. The dataset comprises 10 concepts, each accompanied by 5 single-concept images, as illustrated in Figure 8 along with their corresponding class names. For multi-concept images, we randomly grouped the concepts into 5 pairs, collecting 5 images for each pair. Additionally, we employed a pre-trained detection model, Grounding DINO (Liu et al., 2023), to identify the locations of each concept in the input images, generating corresponding bounding boxes. These bounding boxes served as input for a pre-trained segmentation model, SAM (Kirillov et al., 2023), which produced segmentation maps for each concept. Samples of the multi-concept images and the segmentation maps for individual concepts are displayed in Figure 9.

| pot | penbag | helmet | headphone | bucket |
| doll | robot toy | dinosaur toy | cup | cat figurine |

Figure 8. Samples of the single-concept images of each concept in the dataset

B. Complementary Experimental Results

Due to space constraints, only a summary of the experimental results is provided in Section 4. For a more in-depth examination, detailed experimental results and analysis are presented in this section.

B.1. Single-concept generation

The CLIP-I, CLIP-T, KID, and LPIPS scores for our method and the baseline models on each single concept are detailed in Table 3 to Table 6. Notably, *DreamBooth* fine-tuned on single-concept images achieves the highest image fidelity, securing the top CLIP-I score on all single concepts except for the headphone (where it obtains the second-best score) and the lowest KID score across all single concepts. Interestingly, *DreamBooth* fine-tuned on multi-concept images obtains the second-best CLIP-I and KID scores. This suggests that fine-tuning more parameters can be advantageous for learning visual representation from input images. However, a trade-off between acquiring visual representation and losing prior semantic knowledge is evident, as both variants of *DreamBooth* generally yield lower CLIP-T and LPIPS scores compared to other methods. As evident in Figure 3 and Figure 10, *DreamBooth* struggles to fully capture semantic information from text prompts.

Comparatively, when assessing our method against *Custom Diffusion*, both the hard guidance variant and soft guidance variant outperform *Custom Diffusion* fine-tuned on multi-concept images across all evaluation metrics. This superiority can be attributed to *Custom Diffusion* having a tendency to generate all concepts from multi-concept images, impacting the image fidelity and image-text alignment of the generated images. Examples of generating all input concepts are illustrated in

Figure 9. Samples of the multi-concept images with segmentation maps of individual concepts

Figure 3 and Figure 10. In contrast, our method can precisely specify the target concept, resulting in improved performance. While the CLIP-I and KID scores of our method closely align with those of *Custom Diffusion* fine-tuned on single-concept images, the CLIP-T and LPIPS scores are higher.

B.2. Multi-concept generation

The CLIP-I, CLIP-T, KID, and LPIPS scores for our method and the baseline models on multi-concept groups are detailed in Table 3 to Table 6. In multi-concept generation, the hard guidance variant of our method achieves superior image fidelity compared to the baseline models in three groups. Notably, the hard guidance variant outperforms the soft guidance variant on all multi-concept groups, suggesting that hard guidance can introduce more visual representation to the model. The cross-attention map in Figure 5 further illustrates that only the hard guidance variant aligns with the position of the target concepts in generated images, whereas the soft guidance variant outlines only a part of the target concept, and the attention distribution does not correspond to the respective concepts in multi-concept generation. Hence, we conclude that hard guidance is favorable for multi-concept generation.

Additionally, *Custom Diffusion* trained on single-concept images achieves the highest CLIP-T score, but its image fidelity is low. As evident in Figure 4 and Figure 11, properties of the two input concepts are infused in the generated images. This property infusion is also observed in the images generated by *DreamBooth* fine-tuned on single-concept images. It's worth noting that property infusion does not always occur when jointly training on single-concept images, as evidenced by some successful cases presented in Figure 11. However, for better results in multi-concept generation, we recommend training on multi-concept images using our method, which avoids property infusion and enhances generation quality.

B.3. Ablation study

In the ablation study, we explore the performance of our method when optimizing different parameter sets to identify the optimal trainable parameter set. Specifically, we test three parameter sets from the cross-attention layers in the UNet: $W_Q + W_K + W_V$, $W_Q + W_V$, and $W_K + W_V$. In the cross-attention layers, the W_Q matrix handles information from the image, while the W_K and W_V matrices deal with information from text prompts. Moreover, the W_Q and W_K matrices are involved in the calculation of cross-attention maps, while the W_V matrix carries the most semantic information from text prompts. Consequently, the rates of weight change of the W_V matrix are most significant, as shown in Figure 6. The values

Table 3. Comparison of CLIP-I value between our method and baseline models on each concept in the dataset (Bold indicates the best value, underline represents the second-best value)

Target Concept	Multi-concept Input				Single-concept Input	
	Ours (hard guidance)	Ours (soft guidance)	Custom Diffusion	DreamBooth	Custom Diffusion	DreamBooth
pot	0.4191	0.4114	0.4355	<u>0.6192</u>	0.5002	**0.7431**
penbag	0.4839	0.4852	0.4673	<u>0.6123</u>	0.4950	**0.6522**
helmet	0.5619	0.5514	0.5711	<u>0.6598</u>	0.6037	**0.6743**
headphone	0.6783	0.6347	0.5229	**0.7290**	0.6954	<u>0.6901</u>
bucket	0.4420	0.4529	0.3990	<u>0.5287</u>	0.4618	**0.5657**
doll	0.5012	0.4575	0.4421	<u>0.6091</u>	0.5456	**0.6979**
robot toy	0.5167	0.5285	0.5073	<u>0.5148</u>	0.5558	**0.5905**
dinosaur toy	0.6433	0.6257	0.6010	<u>0.7060</u>	0.6470	**0.7270**
cup	0.4210	0.4059	0.4187	<u>0.4677</u>	0.4601	**0.6270**
cat figurine	0.5592	0.5234	0.5181	<u>0.5531</u>	0.5608	**0.6122**
pot & penbag	0.4548	0.4504	0.4538	**0.5913**	0.4182	<u>0.5189</u>
helmet & headphone	0.4964	0.4493	0.4729	**0.5500**	0.4296	<u>0.5227</u>
bucket & doll	**0.6006**	0.5557	<u>0.5791</u>	0.5236	0.5072	0.5556
robot toy & dinosaur toy	**0.5583**	0.5201	0.5174	0.3931	<u>0.5222</u>	0.5150
cup & cat figurine	**0.5427**	<u>0.4999</u>	0.4304	0.4590	0.4285	0.4776

Table 4. Comparison of CLIP-T value between our method and baseline models on each concept in the dataset (Bold indicates the best value, underline represents the second-best value)

Target Concept	Multi-concept Input				Single-concept Input	
	Ours (hard guidance)	Ours (soft guidance)	Custom Diffusion	DreamBooth	Custom Diffusion	DreamBooth
pot	**0.2354**	0.2328	0.2267	0.1611	<u>0.2341</u>	0.1498
penbag	0.2478	**0.2628**	0.2354	0.2054	<u>0.2492</u>	0.2191
helmet	<u>0.2795</u>	**0.2800**	0.2752	0.2175	0.2617	0.2214
headphone	<u>0.2658</u>	**0.2675**	0.2519	0.2266	0.2612	0.2319
bucket	0.2798	**0.2827**	0.2781	0.2504	<u>0.2815</u>	0.2583
doll	<u>0.2413</u>	**0.2532**	0.2398	0.1591	0.2346	0.1549
robot toy	<u>0.2921</u>	0.2884	0.2904	0.2230	**0.2923**	0.2615
dinosaur toy	<u>0.3008</u>	0.2991	0.2957	0.2679	**0.3066**	0.2614
cup	<u>0.2523</u>	0.2475	0.2449	0.2129	**0.2540**	0.2034
cat figurine	**0.2748**	0.2658	0.2745	0.1754	<u>0.2700</u>	0.1993
pot & penbag	<u>0.2451</u>	0.2369	0.2265	0.1913	**0.2684**	0.2106
helmet & headphone	<u>0.2765</u>	0.2697	0.2713	0.2062	**0.2918**	0.1570
bucket & doll	<u>0.2621</u>	0.2604	0.2551	0.1588	**0.2671**	0.1939
robot toy & dinosaur toy	<u>0.2847</u>	0.2796	0.2808	0.2300	**0.3074**	0.2269
cup & cat figurine	0.2803	0.2726	0.2407	<u>0.2819</u>	**0.3106**	0.1970

of CLIP-I, CLIP-T, KID, and LPIPS metrics for each experiment are presented in Table 7 and Table 8.

The detailed results reveal that optimizing $W_Q + W_K + W_V$ is advantageous for achieving better image fidelity, as it obtains the best CLIP-I and KID scores for most target concepts. However, its performance on CLIP-T and LPIPS is inferior to the other sets. This observation aligns with the comparison between *DreamBooth* and *Custom Diffusion*, highlighting the trade-off between the acquisition of visual representations and the loss of prior semantic knowledge. Furthermore, there is no significant theoretical distinction between applying $W_Q + W_V$ or $W_K + W_V$. However, based on the experimental results, $W_K + W_V$ yields better performance than $W_Q + W_V$. Considering the overall performance across all evaluation metrics, $W_K + W_V$ is deemed the optimal choice for trainable parameters.

Table 5. Comparison of KID value between our method and baseline models on each concept in the dataset (Bold indicates the best value, underline represents the second-best value)

Target Concept	Multi-concept Input				Single-concept Input	
	Ours (hard guidance)	Ours (soft guidance)	Custom Diffusion	DreamBooth	Custom Diffusion	DreamBooth
pot	0.1695	0.2026	0.1609	<u>0.0737</u>	0.1621	**0.0528**
penbag	0.2269	0.2326	0.2118	<u>0.1347</u>	0.1984	**0.0877**
helmet	0.1305	0.1505	0.1328	0.1473	<u>0.1163</u>	**0.0618**
headphone	0.2173	0.2462	0.2885	<u>0.1268</u>	0.1306	**0.1257**
bucket	0.2171	0.2165	0.2337	<u>0.1718</u>	0.1991	**0.1672**
doll	0.1515	0.1565	0.1601	<u>0.1010</u>	0.1270	**0.0724**
robot toy	0.3108	0.3083	0.3124	0.3501	<u>0.2739</u>	**0.2468**
dinosaur toy	0.1554	0.1410	0.1709	<u>0.1200</u>	0.1505	**0.0964**
cup	0.0895	0.1105	0.1029	0.1152	<u>0.0954</u>	**0.0537**
cat figurine	0.4239	0.4406	0.4544	0.6443	<u>0.4120</u>	**0.3320**
pot & penbag	0.2259	0.2541	<u>0.2170</u>	**0.1195**	0.3051	0.2612
helmet & headphone	0.1174	0.1155	<u>0.1140</u>	**0.1098**	0.1666	0.1132
bucket & doll	0.1088	<u>0.1028</u>	0.1125	**0.0890**	0.1436	0.1370
robot toy & dinosaur toy	**0.2484**	0.3074	<u>0.2942</u>	0.3788	0.3302	0.3389
cup & cat figurine	**0.1039**	<u>0.1110</u>	0.1339	0.1156	0.1438	0.1255

Table 6. Comparison of LPIPS value between our method and baseline models on each concept in the dataset (Bold indicates the best value, underline represents the second-best value)

Target Concept	Multi-concept Input				Single-concept Input	
	Ours (hard guidance)	Ours (soft guidance)	Custom Diffusion	DreamBooth	Custom Diffusion	DreamBooth
pot	0.6578	**0.6702**	<u>0.6697</u>	0.6059	0.6456	0.4382
penbag	**0.6669**	0.6631	<u>0.6659</u>	0.6598	0.6488	0.6229
helmet	**0.6651**	<u>0.6544</u>	0.6543	0.6021	0.6313	0.6505
headphone	0.6172	<u>0.6266</u>	**0.6349**	0.5825	0.5583	0.4841
bucket	**0.6968**	<u>0.6909</u>	0.6874	0.6894	0.6669	0.6407
doll	<u>0.7006</u>	**0.7037**	0.6896	0.6819	0.6776	0.6596
robot toy	<u>0.6802</u>	**0.6833**	0.6561	0.6085	0.6336	0.5734
dinosaur toy	0.6130	**0.6338**	0.6109	<u>0.6308</u>	0.6278	0.6022
cup	<u>0.6805</u>	**0.6841**	0.6803	0.6186	0.6572	0.5958
cat figurine	**0.6767**	<u>0.6744</u>	0.6461	0.4825	0.6084	0.6169
pot & penbag	**0.6672**	<u>0.6643</u>	0.6459	0.5514	0.6523	0.6613
helmet & headphone	0.6662	<u>0.6710</u>	0.6636	0.5854	**0.6780**	0.6566
bucket & doll	0.5816	0.6204	0.5941	**0.6301**	<u>0.6268</u>	0.5292
robot toy & dinosaur toy	0.6242	<u>0.6448</u>	0.6162	**0.6464**	0.6318	0.6258
cup & cat figurine	0.6268	**0.6539**	0.4252	0.6048	<u>0.6536</u>	0.5129

Figure 10. Complementary result of qualitative comparison in single-concept generation

Figure 11. Complementary result of qualitative comparison in multi-concept generation

Table 7. Comparison of CLIP-I and CLIP-T value between different parameter sets for optimization (Bold indicates the best value, underline represents the second-best value)

Target Concept	CLIP-I ↑			CLIP-T ↑		
	$W_Q + W_K + W_V$	$W_Q + W_V$	$W_K + W_V$	$W_Q + W_K + W_V$	$W_Q + W_V$	$W_K + W_V$
pot	**0.4313**	<u>0.4201</u>	0.4191	<u>0.2345</u>	0.2335	**0.2354**
penbag	<u>0.4930</u>	**0.4934**	0.4839	<u>0.2405</u>	**0.2523**	<u>0.2478</u>
helmet	<u>0.5573</u>	0.5557	**0.5619**	0.2722	<u>0.2791</u>	**0.2795**
headphone	<u>0.6674</u>	0.6483	**0.6783**	0.2626	<u>0.2629</u>	**0.2658**
bucket	<u>0.4400</u>	0.4398	**0.4420**	0.2780	<u>0.2797</u>	**0.2798**
doll	**0.5149**	0.4989	<u>0.5012</u>	**0.2447**	<u>0.2441</u>	0.2413
robot toy	<u>0.5247</u>	**0.5345**	0.5167	**0.2982**	<u>0.2951</u>	0.2921
dinosaur toy	**0.6495**	0.6410	<u>0.6433</u>	0.2948	<u>0.2999</u>	**0.3008**
cup	**0.4365**	0.4010	<u>0.4210</u>	0.2428	<u>0.2479</u>	**0.2523**
cat figurine	**0.5732**	0.5091	<u>0.5592</u>	0.2602	<u>0.2639</u>	**0.2748**
pot & penbag	<u>0.4733</u>	**0.4742**	0.4548	<u>0.2380</u>	0.2331	**0.2451**
helmet & headphone	**0.4981**	0.4626	<u>0.4964</u>	<u>0.2762</u>	0.2739	**0.2765**
bucket & doll	**0.6135**	0.5760	<u>0.6006</u>	0.2600	<u>0.2560</u>	**0.2621**
robot toy & dinosaur toy	0.5346	<u>0.5513</u>	**0.5583**	0.2821	**0.2869**	<u>0.2847</u>
cup & cat figurine	0.4924	<u>0.5019</u>	**0.5427**	0.2659	<u>0.2794</u>	**0.2803**

Table 8. Comparison of KID and LPIPS value between different parameter sets for optimization (Bold indicates the best value, underline represents the second-best value)

Target Concept	KID ↓			LPIPS ↑		
	$W_Q + W_K + W_V$	$W_Q + W_V$	$W_K + W_V$	$W_Q + W_K + W_V$	$W_Q + W_V$	$W_K + W_V$
pot	**0.1461**	<u>0.1570</u>	0.1695	**0.6659**	0.6517	<u>0.6578</u>
penbag	**0.1675**	<u>0.2091</u>	0.2269	<u>0.6645</u>	0.6616	**0.6669**
helmet	**0.1292**	0.1355	<u>0.1305</u>	<u>0.6649</u>	0.6606	**0.6651**
headphone	**0.2018**	<u>0.2094</u>	0.2173	0.6089	**0.6218**	<u>0.6172</u>
bucket	<u>0.2208</u>	0.2269	**0.2171**	<u>0.6874</u>	0.6871	**0.6968**
doll	<u>0.1477</u>	**0.1424**	0.1515	0.6954	**0.7100**	<u>0.7006</u>
robot toy	<u>0.2982</u>	**0.2940**	0.3108	<u>0.6768</u>	0.6747	**0.6802**
dinosaur toy	<u>0.1720</u>	0.1799	**0.1554**	<u>0.6087</u>	0.6086	**0.6130**
cup	**0.0892**	0.0988	<u>0.0895</u>	0.6592	<u>0.6605</u>	**0.6805**
cat figurine	**0.4128**	0.4508	<u>0.4239</u>	0.6550	<u>0.6646</u>	**0.6767**
pot & penbag	**0.1823**	<u>0.1967</u>	0.2259	0.6610	<u>0.6593</u>	**0.6672**
helmet & headphone	<u>0.1120</u>	**0.1112**	0.1174	**0.6734**	<u>0.6728</u>	0.6662
bucket & doll	**0.1068**	<u>0.1090</u>	0.1088	0.5637	<u>0.5738</u>	**0.5816**
robot toy & dinosaur toy	<u>0.2478</u>	0.2459	**0.2484**	**0.6271**	0.6214	<u>0.6242</u>
cup & cat figurine	<u>0.1075</u>	0.1289	**0.1039**	**0.6314**	0.6244	<u>0.6268</u>