# *RealCustom*: Narrowing Real Text Word for Real-Time Open-Domain Text-to-Image Customization

Mengqi Huang[1]*, Zhendong Mao[1]†, Mingcong Liu[2], Qian He[2], Yongdong Zhang[1]
[1] University of Science and Technology of China; [2]ByteDance Inc.
{huangmq}@mail.ustc.edu.cn, {zdmao, zhyd73}@ustc.edu.cn, {liumingcong, heqian}@bytedance.com

## Abstract

*Text-to-image customization, which aims to synthesize text-driven images for the given subjects, has recently revolutionized content creation. Existing works follow the pseudo-word paradigm, i.e., represent the given subjects as pseudo-words and then compose them with the given text. However, the inherent entangled influence scope of pseudo-words with the given text results in a dual-optimum paradox, i.e., the similarity of the given subjects and the controllability of the given text could not be optimal simultaneously. We present* ***RealCustom*** *that, for the first time, disentangles similarity from controllability by precisely limiting subject influence to relevant parts only, achieved by gradually narrowing* ***real*** *text word from its general connotation to the specific subject and using its cross-attention to distinguish relevance. Specifically, RealCustom introduces a novel "train-inference" decoupled framework: (1) during training, RealCustom learns general alignment between visual conditions to original textual conditions by a novel adaptive scoring module to adaptively modulate influence quantity; (2) during inference, a novel adaptive mask guidance strategy is proposed to iteratively update the influence scope and influence quantity of the given subjects to gradually narrow the generation of the real text word. Comprehensive experiments demonstrate the superior real-time customization ability of RealCustom in the open domain, achieving both unprecedented similarity of the given subjects and controllability of the given text for the first time. The project page is* https://corleone-huang.github.io/realcustom/.

## 1. Introduction

Recent significant advances in the customization of pre-trained large-scale text-to-image models [6, 24, 25, 28] (*i.e.*, *text-to-image customization*) has revolutionized content creation. This task empowers pre-trained models with the ability to generate imaginative text-driven scenes for subjects specified by users, which is a foundation for AI-generated content (AIGC) and real-world applications such as personal image and video creation [7]. The primary goal of customization is dual-faceted: (1) high-quality *similarity*, *i.e.*, the target subjects in the generated images should closely mirror the *given subjects*; (2) high-quality *controllability*, *i.e.*, the remaining subject-irrelevant parts should consistently adhere to the control of the *given text*.

Existing literature follows the *pseudo-word* paradigm, *i.e.*, (1) learning pseudo-words (*e.g.*, $S^*$ [10] or rare-tokens [27]) to represent the given subjects; (2) composing these

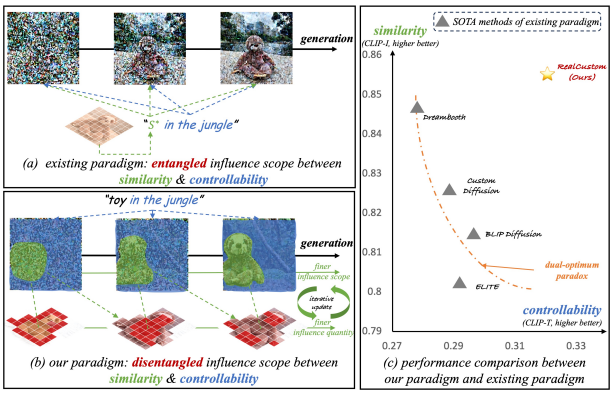


Figure 1. Comparison between the existing paradigm and ours. **(a)** The existing paradigm represents the *given subject* as pseudo-words (*e.g.*, $S^*$), which has entangled the same entire influence scope with the *given text*, resulting in the *dual-optimum paradox*, *i.e.*, the similarity for the *given subject* and the controllability for the *given text* could not achieve optimum simultaneously. **(b)** We propose *RealCustom*, for the first time disentangles similarity from controllability by precisely limiting the *given subjects* to influence only the relevant parts while the rest parts are purely controlled by the *given text*. This is achieved by iteratively updating the influence scope and influence quantity of the *given subjects*. **(c)** The quantitative comparison shows that our paradigm achieves both superior similarity and controllability to the existing paradigm.

*Works done during the internship at ByteDance.
†Zhendong Mao is the corresponding author.


Figure 2. Generated customization results of our proposed novel paradigm ***RealCustom***. Given a *single* image representing the given subject in the open domain (***any subjects***, portrait painting, favorite toys, *etc*.), *RealCustom* could generate realistic images that consistently adhere to the given text for the given subjects in real-time (***without any test-time optimization steps***).

pseudo-words with the given text for the customized generation. Recent studies have focused on learning more comprehensive pseudo-words [1, 8, 22, 32, 38] to capture more subject information, *e.g*., different pseudo-words for different diffusion timesteps [1, 38] or layers [32]. Meanwhile, others propose to speed up pseudo-word learning by training an encoder [11, 18, 30, 34] on object-datasets [17]. In parallel, based on the learned pseudo-words, many works further finetune the pre-trained models [16, 18, 27, 34] or add additional adapters [30] for higher similarity. As more information of the given subjects is introduced into pre-trained models, the risk of overfitting increases, leading to the degradation of controllability. Therefore, various regularizations (*e.g*., $l_1$ penalty [10, 16, 34], prior-preservation loss [27]) are used to maintain controllability, which in turn sacrifices similarity. *Essentially*, existing methods are trapped in a *dual-optimum paradox*, *i.e*., the similarity and controllability can not be optimal simultaneously.

We argue that the fundamental cause of this *dual-optimum paradox* is rooted in the existing pseudo-word paradigm, where the similarity component (*i.e*., the pseudo-words) to generate the given subjects is intrinsically *entangled* with the controllability component (*i.e*., the given text) to generate subject-irrelevant parts, causing an overall conflict in the generation, as illustrated in Fig. 1(a). Specifically, this entanglement is manifested in the same entire influence scope of these two components. *i.e*., both the pseudo-words and the given text affect all generation regions. This is because each region is updated as a weighted sum of all word features through built-in textual cross-attention in pre-trained text-to-image diffusion models. Therefore, increasing the influence of the similarity component will simultaneously strengthen the similarity in the subject-relevant parts and weaken the influence of the given text in other irrelevant ones, causing the degradation of controllability, and *vice versa*. Moreover, the necessary correspondence between pseudo-words and subjects confines existing methods to either lengthy test-time optimization [10, 16, 27] or training [18, 34] on object-datasets [17] that have limited categories. As a result, the existing paradigm inherently has poor generalization capability for real-time open-domain scenarios in the real world.

In this paper, we present ***RealCustom***, a novel customization paradigm that, for the first time, disentangles the similarity component from the controllability component by precisely limiting the given subjects to influence only the relevant parts while maintaining other irreverent ones purely controlled by the given texts, achieving both high-quality similarity and controllability in a real-time open-domain scenario, as shown in Fig. 2. The core idea of *RealCustom* is that, instead of representing subjects as pseudo-words, we could progressively narrow down the ***real*** text words (*e.g*., "toy") from their initial general connotation (*e.g*., various kinds o toys) to the specific subjects (*e.g*., the unique sloth toy), wherein the superior text-image alignment in pre-trained models' cross-attention can be leveraged to distinguish subject relevance, as illustrated in Fig. 1(b). Specifically, at each generation step, (1) the influence scope of the given subject is identified by the target real word's cross-attention, with a higher attention score indicating greater relevance; (2) this influence scope then determines the influence quantity of the given subject at the current step, *i.e*., the amount of subject information to be infused into this scope; (3) this influence quantity, in turn, shapes a more accurate influence scope for the next step, as each step's generation result is based on the output of the previous. Through this iterative updating, the generation result of the real word is smoothly and accurately transformed into the given subject, while other irrelevant parts are completely controlled by the given text.

Technically, *RealCustom* introduces an innovative "train-inference" decoupled framework: (1) During training, *RealCustom* only learns the generalized alignment capabilities between visual conditions and pre-trained models' original text conditions on large-scale text-image datasets through a novel *adaptive scoring module*, which modulates the influence quantity based on text and currently generated features. (2) During inference, real-time customization is achieved by a novel *adaptive mask guidance strategy*, which gradually narrows down a real text word based on the learned alignment capabilities. Specifically, (1) the *adaptive scoring module* first estimates the visual features' correlation scores with the text features and currently generated features, respectively. Then a timestep-aware schedule is applied to fuse these two scores. A subset of key visual features, chosen based on the fused score, is

incorporated into pre-trained diffusion models by extending its textual cross-attention with another visual cross-attention. (2) The *adaptive mask guidance strategy* consists of a *text-to-image (T2I)* branch (with the visual condition set to $\mathbf{0}$) and a *text&image-to-image (TI2I)* branch (with the visual condition set to the given subject). Firstly, all layers' cross-attention maps of the target real word in the T2I branch are aggregated into a single one, selecting only high-attention regions as the influence scope. Secondly, in the TI2I branch, the influence scope is multiplied by currently generated features to produce the influence quantity and concurrently multiplied by the outputs of the visual cross-attention to avoid influencing subject-irrelevant parts.

Our contributions are summarized as follows:

Concepts. For the first time, we (1) point out the *dual-optimum paradox* is rooted in the existing pseudo-word paradigm's entangled influence scope between the similarity (*i.e.*, pseudo-words representing the given subjects) and controllability (*i.e.*, the given texts); (2) present *RealCustom*, a novel paradigm that achieves disentanglement by gradually narrowing down *real* words into the given subjects, wherein the given subjects' influence scope is limited based on the cross-attention of the real words.

Technology. The proposed *RealCustom* introduces a novel "train-inference" decoupled framework: (1) during training, learning generalized alignment between visual conditions to original text conditions by the *adaptive scoring module* to modulate influence quantity; (2) during inference, the *adaptive mask guidance strategy* is proposed to narrow down a real word by iterative updating the given subject's influence scope and quantity.

Significance. For the first time, we achieve (1) superior similarity and controllability *simultaneously*, as shown in Fig. 1(c); (2) real-time open-domain customization ability.

## 2. Related Works

### 2.1. Text-to-Image Customization

Existing customization methods follow the *pseudo-words* paradigm, *i.e.*, representing the given subjects as *pseudo-words* and then composing them with the given text for customization. Since the necessary correspondence between the pseudo-words and the given subjects, existing works are confined to either cumbersome test-time optimization-based [1, 8–10, 16, 22, 27, 32] or encoder-based [7, 11, 14, 18, 30, 34] that trained on object-datasets with limited categories. For example, in the optimization-based stream, DreamBooth [27] uses a rare-token as the pseudo-word and further fine-tunes the entire pre-trained diffusion model for better similarity. Custom Diffusion [16] instead finds a subset of key parameters and only optimizes them. The main drawback of this stream is that it requires lengthy optimization times for each new subject. As for the encoder-based stream, the recent ELITE [34] uses a local mapping network to improve similarity, while BLIP-Diffusion [18] introduces a multimodal encoder for better subject representation. These encoder-based works usually show less similarity than optimization-based works and generalize poorly to unseen categories in training. *In summary*, the entangled influence scope of pseudo-words and the given text naturally limits the current works from achieving both optimal similarity and controllability, as well as hindering real-time open-domain customization.

### 2.2. Cross-Attention in Diffusion Models

Text guidance in modern large-scale text-to-image diffusion models [2, 6, 24, 25, 28] is generally performed using the cross-attention mechanism. Therefore, many works propose to manipulate the cross-attention map for text-driven editing [3, 12] on generated images or real images via inversion [31], *e.g.*, Prompt-to-Prompt [12] proposes to reassign the cross-attention weight to edit the generated image. Another branch of work focuses on improving cross-attention either by adding additional spatial control [20, 21] or post-processing to improve semantic alignment [5, 19]. Meanwhile, a number of works [33, 35, 36] propose using cross-attention in diffusion models for discriminative tasks such as segmentation. However, different from the existing literature, the core idea of *RealCustom* is to gradually narrow a real text word from its initial general connotation (*e.g.*, whose cross-attention could represent any toy with various types of shapes and details) to the unique given subject (*e.g.*, whose cross-attention accurately represents the unique toy), which is completely unexplored.

## 3. Methodology

In this study, we focus on the most general customization scenario: with only a *single* image representing the given subject, generating new high-quality images for that subject from the given text. The generated subject may vary in location, pose, style, *etc.*, yet it should maintain high *similarity* with the given one. The remaining parts should consistently adhere to the given text, thus ensuring *controllability*.

We first briefly introduce the preliminaries in Sec. 3.1. The training and inference paradigm of *RealCustom* will be elaborated in detail in Sec. 3.2 and Sec. 3.3, respectively.

### 3.1. Preliminaries

Our paradigm is implemented over Stable Diffusion [25], which consists of two components, *i.e.*, an autoencoder and a conditional UNet [26] denoiser. Firstly, given an image $\boldsymbol{x} \in \mathbb{R}^{H\times W\times 3}$, the encoder $\mathcal{E}(\cdot)$ of the autoencoder maps it into a lower dimensional latent space as $\boldsymbol{z} = \mathcal{E}(\boldsymbol{x}) \in \mathbb{R}^{h\times w\times c}$, where $f = \frac{H_0}{h} = \frac{W_0}{w}$ is the downsampling factor and $c$ stands for the latent channel dimension. The corresponding decoder $\mathcal{D}(\cdot)$ maps the latent vectors back to the

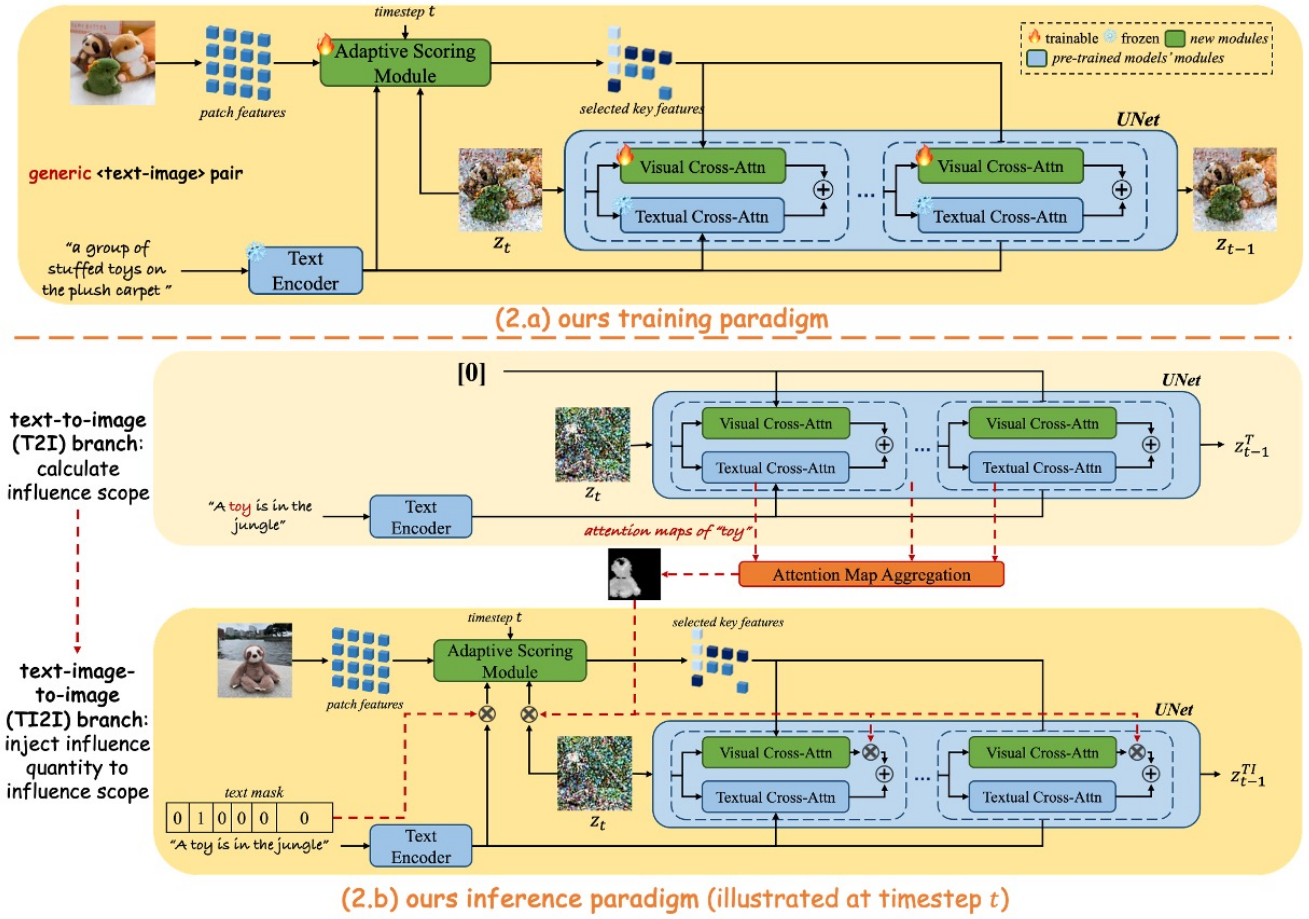


Figure 3. Illustration of our proposed *RealCustom*, which employs a novel "train-inference" decoupled framework: (a) During training, general alignment between visual and original text conditions is learned by the proposed *adaptive scoring module*, which accurately derives visual conditions based on text and currently generated features. (b) During inference, progressively narrowing down a real word (*e.g.*, "toy") from its initial general connotation to the given subject (*e.g.*, the unique brown sloth toy) by the proposed *adaptive mask guidance strategy*, which consists of two branches, *i.e.*, a text-to-image (T2I) branch where the visual condition is set to $\mathbf{0}$, and a text&image-to-image (TI2I) branch where the visual condition is set to the given subject. The T2I branch aims to calculate the influence scope by aggregating the target real word's (*e.g.*, "toy") cross-attention, while the TI2I branch aims to inject the influence quantity into this scope.

image as $\mathcal{D}(\mathcal{E}(\boldsymbol{x})) \approx \boldsymbol{x}$. Secondly, the conditional denoiser $\epsilon_\theta(\cdot)$ is trained on this latent space to generate latent vectors based on the text condition $y$. The pre-trained CLIP text encoder [23] $\tau_{\text{text}}(\cdot)$ is used to encode the text condition $y$ into text features $\boldsymbol{f_{ct}} = \tau_{\text{text}}(y)$. Then, the denoiser is trained with mean-squared loss:

$$L := \mathbb{E}_{\boldsymbol{z}\sim\mathcal{E}(\boldsymbol{x}),\boldsymbol{f_y},\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathrm{I}),t}\left[\|\boldsymbol{\epsilon} - \epsilon_\theta\left(\boldsymbol{z_t}, t, \boldsymbol{f_{ct}}\right)\|_2^2\right], \quad (1)$$

where $\boldsymbol{\epsilon}$ denotes for the unscaled noise and $t$ is the timestep. $\boldsymbol{z_t}$ is the latent vector that noised according to $t$:

$$\boldsymbol{z_t} = \sqrt{\hat{\alpha}_t}\boldsymbol{z_0} + \sqrt{1-\hat{\alpha}_t}\boldsymbol{\epsilon}, \quad (2)$$

where $\hat{\alpha}_t \in [0, 1]$ is the hyper-parameter that modulates the quantity of noise added. Larger $t$ means smaller $\hat{\alpha}_t$ and thereby a more noised latent vector $\boldsymbol{z_t}$. During inference, a random Gaussian noise $\boldsymbol{z_T}$ is iteratively denoised to $\boldsymbol{z_0}$, and the final generated image is obtained through $\boldsymbol{x}' = \mathcal{D}(\boldsymbol{z_0})$.

The incorporation of text condition in Stable Diffusion is implemented as textual cross-attention:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d}})\boldsymbol{V}, \quad (3)$$

where the query $\boldsymbol{Q} = \boldsymbol{W_Q} \cdot \boldsymbol{f_i}$, key $\boldsymbol{K} = \boldsymbol{W_K} \cdot \boldsymbol{f_{ct}}$ and value $\boldsymbol{V} = \boldsymbol{W_V} \cdot \boldsymbol{f_{ct}}$. $\boldsymbol{W_Q}, \boldsymbol{W_K}, \boldsymbol{W_V}$ are weight parameters of query, key and value projection layers. $\boldsymbol{f_i}, \boldsymbol{f_{ct}}$ are the latent image features and text features, and $d$ is the channel dimension of key and query features. The latent image feature is then updated with the attention block output.

### 3.2. Training Paradigm

As depicted in Fig. 3(a), the text $y$ and image $x$ are first encoded into text features $\boldsymbol{f_{ct}} \in \mathbb{R}^{n_t \times c_t}$ and image features $\boldsymbol{f_{ci}} \in \mathbb{R}^{n_i \times c_i}$ by the pre-trained CLIP text/image encoders [23] respectively. Here, $n_t, c_t, n_i, c_i$ are text feature number/dimension and image feature number/dimension, respectively. Afterward, the *adaptive scoring module* takes the text features $\boldsymbol{f_{ct}}$, currently generated features $\boldsymbol{z_t} \in \mathbb{R}^{h \times w \times c}$, and timestep $t$ as inputs to estimate the score for each features in $\boldsymbol{f_{ci}}$, selecting a subset of key ones as the visual condition $\hat{\boldsymbol{f}}_{\boldsymbol{ci}} \in \mathbb{R}^{\hat{n}_i \times c_i}$, where $\hat{n}_i < n_i$ is the selected image feature number. Next, we extend textual cross-attention with another visual cross-attention to incorporate the visual condition $\hat{\boldsymbol{f}}_{\boldsymbol{yi}}$. Specifically, Eq. 3 is rewritten as:

$$\begin{aligned}&\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}, \boldsymbol{K_i}, \boldsymbol{V_i}) = \\ &\text{Softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d}})\boldsymbol{V} + \text{Softmax}(\frac{\boldsymbol{Q}\boldsymbol{K_i}^\top}{\sqrt{d}})\boldsymbol{V_i}, \quad (4)\end{aligned}$$

where the new key $\boldsymbol{K_i} = \boldsymbol{W_{Ki}} \cdot \hat{\boldsymbol{f}}_{\boldsymbol{ci}}$, value $\boldsymbol{V_i} = \boldsymbol{W_{Vi}} \cdot \hat{\boldsymbol{f}}_{\boldsymbol{ci}}$ are added. $\boldsymbol{W_{Ki}}$ and $\boldsymbol{W_{Vi}}$ are weight parameters. During training, only the *adaptive scoring module* and projection layers $\boldsymbol{W_{Ki}}, \boldsymbol{W_{Vi}}$ in each attention block are trainable, while other pre-trained models' weight remains frozen.

**Adaptive Scoring Module**. Unlike the training stage, where the same images as the visual conditions and inputs to the denoiser $\epsilon_\theta$, the given subjects, and the inference generation results should maintain similarity only in the subject part. Therefore, utilizing all image features as visual conditions results in a "train-inference" gap, which will degrade both similarity and controllability at inference.

The above rationale motivates the *adaptive scoring module*, which provides smooth and accurate visual conditions for customization. As illustrated in Fig. 4, the text $\boldsymbol{f_{ct}} \in \mathbb{R}^{n_t \times c_t}$ and currently generated features $\boldsymbol{z_t} \in \mathbb{R}^{h \times w \times c} = \mathbb{R}^{n_z \times c}$ are first aggregated into the textual context $\boldsymbol{C}_{\text{textual}}$ and visual context $\boldsymbol{C}_{\text{visual}}$ through weighted pooling:

$$\boldsymbol{A}_{\text{textual}} = \text{Softmax}(\boldsymbol{f_{ct}}\boldsymbol{W_a^t}) \in \mathbb{R}^{n_t \times 1} \quad (5)$$

$$\boldsymbol{A}_{\text{visual}} = \text{Softmax}(\boldsymbol{z_t}\boldsymbol{W_a^v}) \in \mathbb{R}^{n_z \times 1} \quad (6)$$

$$\boldsymbol{C}_{\text{textual}} = \boldsymbol{A}_{\text{textual}}^\top \boldsymbol{f_{ct}} \in \mathbb{R}^{1 \times c_t}, \boldsymbol{C}_{\text{visual}} = \boldsymbol{A}_{\text{visual}}^\top \boldsymbol{z_t} \in \mathbb{R}^{1 \times c}, \quad (7)$$

where $\boldsymbol{W_a^t} \in \mathbb{R}^{c_t \times 1}, \boldsymbol{W_a^v} \in \mathbb{R}^{c \times 1}$ are weight parameters, and "Softmax" is operated in the number dimension. These contexts are then spatially replicated and concatenated with image features $\boldsymbol{f_{ci}} \in \mathbb{R}^{n_i \times c_i}$ to estimate the textual score $\boldsymbol{S}_{\text{textual}} \in \mathbb{R}^{n_i \times 1}$ and visual score $\boldsymbol{S}_{\text{visual}} \in \mathbb{R}^{n_i \times 1}$ respectively. These two scores are predicted by two lightweight score-net, which are implemented as two-layer MLPs.

Considering that the textual features are roughly accurate and the generated features are gradually refined, a timestep-aware schedule is proposed to fuse these two scores:

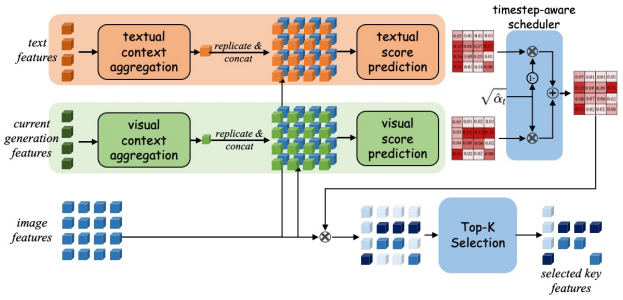


Figure 4. Illustration of *adaptive scoring module*. Text features and currently generated features are first aggregated into the textual and visual context, which are then spatially concatenated with image features to predict textual and visual scores. These scores are then fused based on the current timestep. Ultimately, only a subset of the key features is selected based on the fused score.

$$\boldsymbol{S} = (1 - \sqrt{\hat{\alpha}_t})\boldsymbol{S}_{\text{textual}} + \sqrt{\hat{\alpha}_t}\boldsymbol{S}_{\text{visual}}, \quad (8)$$

where $\sqrt{\hat{\alpha}_t}$ is the hyperparameter of pre-trained diffusion models that modulate the amount of noise added to generated features. Then a softmax activation is applied to the fused score since our focus is on highlighting the comparative significance of each image feature vis-à-vis its counterparts: $\boldsymbol{S} = \text{Softmax}(\boldsymbol{S})$. The fused scores are multiplied with the image features to enable the learning of score-nets:

$$\boldsymbol{f_{ci}} = \boldsymbol{f_{ci}} \circ (1 + \boldsymbol{S}), \quad (9)$$

where $\circ$ denotes the element-wise multiply. Finally, given a Top-K ratio $\gamma_{\text{num}} \in [0, 1]$, a sub-set of key features with highest scores are selected as the output $\hat{\boldsymbol{f}}_{\boldsymbol{yi}} \in \mathbb{R}^{\hat{n}_i \times c_i}$, where $\hat{n}_i = \gamma_{\text{num}} n_i$. To enable flexible inference with different $\gamma_{\text{num}}$ without performance degradation, we propose to use a uniformly random ratio during training:

$$\gamma_{\text{num}} = \text{uniform}[\gamma_{\text{num}}^{\text{low}}, \gamma_{\text{num}}^{\text{high}}], \quad (10)$$

where $\gamma_{\text{num}}^{\text{low}}, \gamma_{\text{num}}^{\text{high}}$ are set to $0.3, 1.0$, respectively.

### 3.3. Inference Paradigm

The inference paradigm of *RealCustom* consists of two branches, *i.e.*, a text-to-image (T2I) branch where the visual input is set to $\mathbf{0}$ and a text&image-to-image (TI2I) branch where the visual input is set to given subjects, as illustrated in Fig. 3(b). These two branches are connected by our proposed *adaptive mask guidance strategy*. Specifically, given previous step's output $\boldsymbol{z_t}$, a pure text conditional denoising process is performed in T2I branch to get the output $\boldsymbol{z_{t-1}^T}$, where all layers cross-attention map of the target real word (*e.g.*, "toy") is extracted and resized to the same resolution (the same as the largest map size, *i.e.*, $64 \times 64$ in

| Methods | *controllability* | | *similarity* | | *efficiency* |
|---|---|---|---|---|---|
| | CLIP-T ↑ | ImageReward ↑ | CLIP-I ↑ | DINO-I ↑ | test-time optimize steps |
| Textual Inversion [10] | 0.2546 | -0.9168 | 0.7603 | 0.5956 | 5000 |
| DreamBooth [27] | 0.2783 | 0.2393 | 0.8466 | 0.7851 | 800 |
| Custom Diffusion [16] | 0.2884 | 0.2558 | 0.8257 | 0.7093 | 500 |
| ELITE [34] | 0.2920 | 0.2690 | 0.8022 | 0.6489 | 0 (real-time) |
| BLIP-Diffusion [18] | 0.2967 | 0.2172 | 0.8145 | 0.6486 | 0 (real-time) |
| ***RealCustom*(ours)** | **0.3204** | **0.8703** | **0.8552** | **0.7865** | **0 (real-time)** |

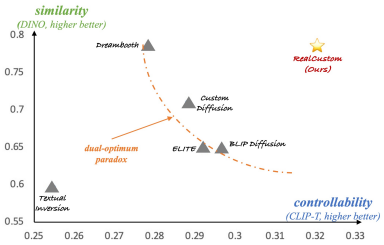


Table 1. Quantitative comparisons. **Left**: Our proposed *RealCustom* outperforms existing methods in all metrics. **Right**: We plot the "CLIP-T verse DINO", showing that the existing methods are trapped into the *dual-optimum paradox*, while *RealCustom* get rid of it and achieve both high-quality similarity and controllability. The same conclusion in "CLIP-T verse CLIP-I" can be found in Fig. 1(c).

Stable Diffusion). The aggregated attention map is denoted as $\boldsymbol{M} \in \mathbb{R}^{64\times64}$. Next, a Top-K selection is applied, *i.e.*, given the target ratio $\gamma_{\text{scope}} \in [0, 1]$, only $\gamma_{\text{scope}} \times 64 \times 64$ regions with the highest cross-attention score will remain, while the rest will be set to 0. The selected cross-attention map $\bar{\boldsymbol{M}}$ is normalized by its maximum value as:

$$\hat{\boldsymbol{M}} = \frac{\bar{\boldsymbol{M}}}{\max(\bar{\boldsymbol{M}})}, \tag{11}$$

where $\max(\cdot)$ represents the maximum value. The rationale behind this is that even in these selected parts, the subject relevance of different regions is also different.

In the TI2I branch, the influence scope $\hat{\boldsymbol{M}}$ is first multiplied by currently generated feature $\boldsymbol{z_t}$ to provide accurate visual conditions for current generation step. The reason is that only subject-relevant parts should be considered for the calculation of influence quantity. Secondly, $\hat{\boldsymbol{M}}$ is multiplied by the visual cross-attention results to prevent negative impacts on the controllability of the given texts in other subject-irrelevant parts. Specifically, Eq. 4 is rewritten as:

$$\begin{aligned}&\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}, \boldsymbol{K_i}, \boldsymbol{V_i}) = \\ &\text{Softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d}})\boldsymbol{V} + (\text{Softmax}(\frac{\boldsymbol{Q}\boldsymbol{K_i}^\top}{\sqrt{d}})\boldsymbol{V_i})\hat{\boldsymbol{M}},\end{aligned} \tag{12}$$

where the necessary resize operation is applied to match the size of $\hat{\boldsymbol{M}}$ with the resolution of each cross-attention block. The denoised output of TI2I branch is denoted as $\boldsymbol{z^{TI}_{t-1}}$. The classifer-free guidance [13] is extended to produce next step's denoised latent feature $\boldsymbol{z_{t-1}}$ as:

$$\boldsymbol{z_{t-1}} = \epsilon_\theta(\emptyset) + \omega_t(\boldsymbol{z^T_{t-1}} - \epsilon_\theta(\emptyset)) + \omega_i(\boldsymbol{z^{TI}_{t-1}} - \boldsymbol{z^T_{t-1}}), \tag{13}$$

where $\epsilon_\theta(\emptyset)$ is the unconditional denoised output.

With the smooth and accurate influence quantity of the given subject injected into the current step, the generation of the real word will gradually be narrowed from its initial general connotation to the specific subject, which will shape a more precise influence scope for the generation of the next step. Through this iterative updating and generation, we achieve real-time customization where the similarity for the given subject is disentangled with the controllability for the given text, leading to an optimal of both. More importantly, since both the *adaptive scoring module* as well as visual cross-attention layers are trained on general text-image datasets, the inference could be generally applied to any categories by using any target real words, enabling excellent open-domain customization capability.

# 4. Experiments

## 4.1. Experimental Setups

**Implementation.** *RealCustom* is implemented on Stable Diffusion and trained on the filtered subset of Laion-5B [29] based on aesthetic score, using 16 A100 GPUs for 16w iterations with 1e-4 learning rate. Unless otherwise specified, DDIM sampler [31] with 50 sample steps is used for sampling and the classifier-free guidance $\omega_t, \omega_i$ is 7.5 and 12.5. Top-K ratios $\gamma_{\text{num}} = 0.8$, $\gamma_{\text{scope}} = 0.25$.

**Evaluation.** ***Similarity.*** We use the state-of-the-art segmentation model (*i.e.*, SAM [15]) to segment the subject, and then evaluate with both CLIP-I and DINO [4] scores, which are average pairwise cosine similarity CLIP ViT-B/32 or DINO embeddings of the segmented subjects in generated and real images. ***Controllability.*** We calculate the cosine similarity between prompt and image CLIP ViT-B/32 embeddings (CLIP-T). In addition, ImageReward [37] is used to evaluate controllability and aesthetics (quality).

**Prior SOTAs.** We compare with existing paradigm of both optimization-based (*i.e.*, Textual Inversion[10], DreamBooth [27], CustomDiffusion [16]) and encoder-based (ELITE[34], BLIP-Diffusion[18]) state-of-the-arts.

## 4.2. Main Results

**Quantitative results.** As shown in Tab. 1, *RealCustom* outperforms existing methods in all metrics: (1) for controllability, we improve CLIP-T and ImageReward by 8.1% and 223.5%, respectively. The significant improvement in ImageReward shows that our paradigm generates much higher

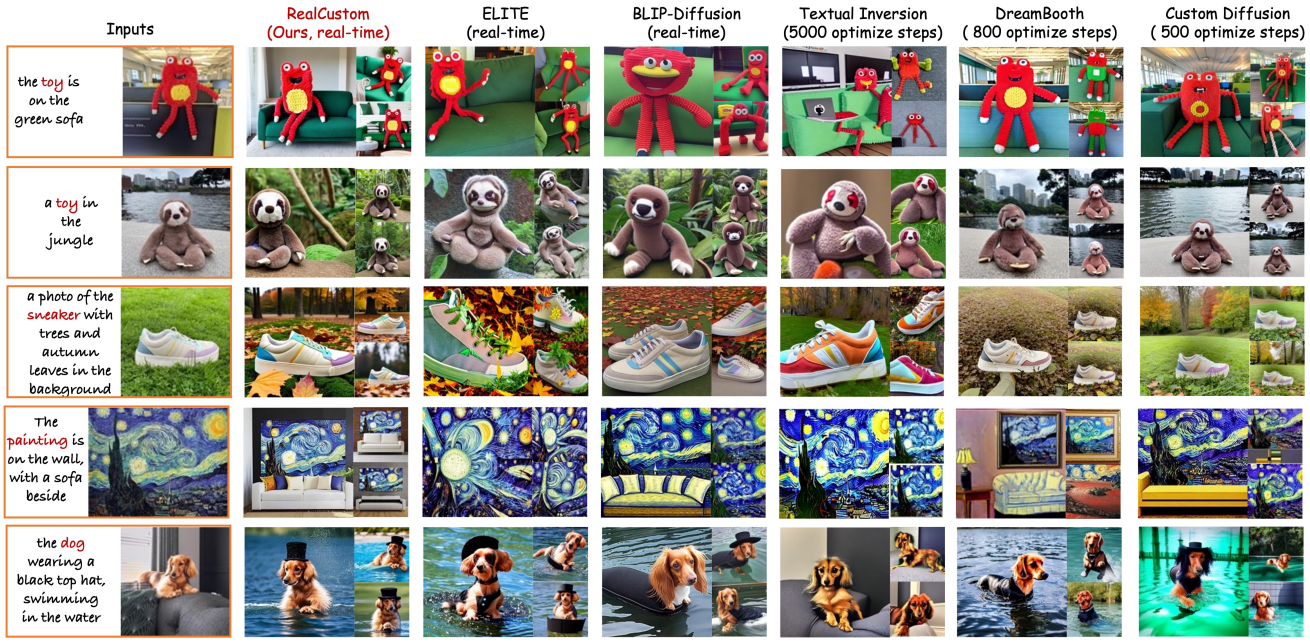


Figure 5. Qualitative comparison with existing methods. *RealCustom* could produce much higher quality customization results that have better similarity with the given subject and better controllability with the given text compared to existing works.

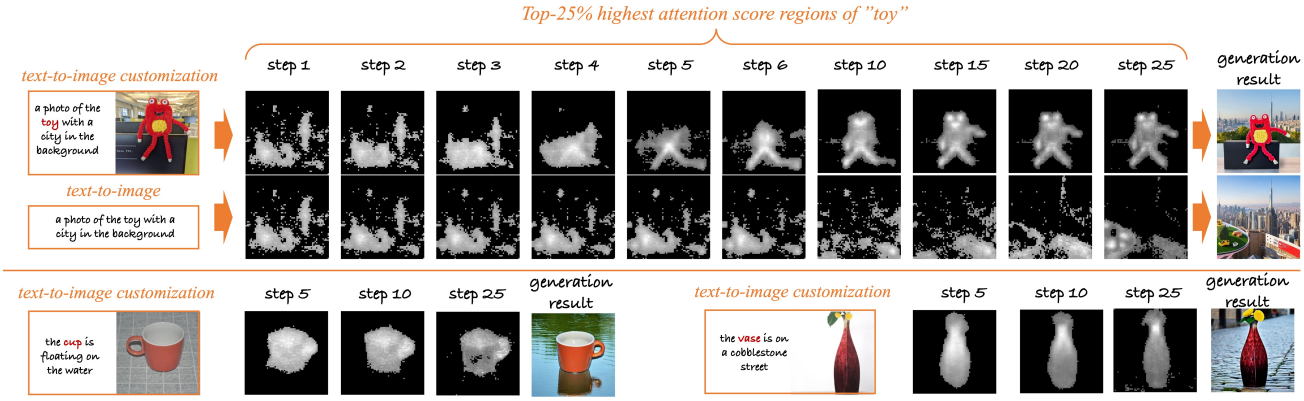


Figure 6. Illustration of gradually narrowing the real words into the given subjects. **Upper**: *RealCustom* generated results (first row) and the original text-to-image generated result (second row) by pre-trained models with the same seed. The mask is visualized by the Top-25% highest attention score regions of the real word "toy". We could observe that starting from the same state (the same mask since there's no information of the given subject is introduced at the beginning), *RealCustom* gradually forms the structure and details of the given subject by our proposed *adaptive mask strategy*, achieving the open-domain zero-shot customization. **Lower**: More visualization cases.

quality customization; (2) for similarity, we also achieve state-of-the-art performance on both CLIP-I and DINO-I. The figure of "CLIP-T verse DINO" validates that the existing paradigm is trapped into the *dual-optimum paradox*, while RealCustom effectively eradicates it.

**Qualitative results.** As shown in Fig. 5, *RealCustom* demonstrates superior zero-shot open-domain customization capability (*e.g.*, the rare shaped toy in the first row), generating higher-quality custom images that have better similarity with the given subject and better controllability with the given text compared to existing works.

### 4.3. Ablations

**Effectiveness of *adaptive mask guidance strategy*.** We first visualize the narrowing down process of the real word by the proposed adaptive mask guidance strategy in Fig. 6. We could observe that starting from the same state (the same mask since there's no information of the given subject is introduced at the first step), *RealCustom* gradually forms the structure and details of the given subject, achiev-

| inference setting | CLIP-T ↑ | CLIP-I ↑ |
|---|---|---|
| $\gamma_{\text{scope}} = 0.1$ | 0.32 | 0.8085 |
| $\gamma_{\text{scope}} = 0.2$ | 0.3195 | 0.8431 |
| $\gamma_{\text{scope}} = 0.25$ | **0.3204** | **0.8552** |
| $\gamma_{\text{scope}} = 0.25$, binary | 0.294 | 0.8567 |
| $\gamma_{\text{scope}} = 0.3$ | 0.3129 | 0.8578 |
| $\gamma_{\text{scope}} = 0.4$ | 0.3023 | 0.8623 |
| $\gamma_{\text{scope}} = 0.5$ | 0.285 | 0.8654 |

Table 2. Ablation of different $\gamma_{\text{scope}}$, which denotes the influence scope of the given subject in *RealCustom* during inference. "binary" means using binary masks instead of max norm in Eq. 11.


Figure 7. Visualization of different influence scope.

| ID | settings | CLIP-T ↑ | CLIP-I ↑ |
|---|---|---|---|
| 1 | full model, $\gamma_{\text{num}} = 0.8$ | **0.3204** | **0.8552** |
| 2 | *w/o* adaptive scoring module | 0.3002 | 0.8221 |
| 3 | textual score only, $\gamma_{\text{num}} = 0.8$ | 0.313 | 0.8335 |
| 4 | visual score only, $\gamma_{\text{num}} = 0.8$ | 0.2898 | 0.802 |
| 5 | (textual + visual) / 2, $\gamma_{\text{num}} = 0.8$ | 0.3156 | 0.8302 |
| 6 | full model, $\gamma_{\text{num}} = 0.9$ | 0.315 | 0.8541 |
| 7 | full model, $\gamma_{\text{num}} = 0.7$ | 0.3202 | 0.8307 |

Table 3. Ablation of the adaptive scoring module, where $\gamma_{\text{num}}$ means the influence quantity of the given subject during inference.

ing the open-domain zero-shot customization while remaining other subject-irrelevant parts (*e.g.*, the city background) completely controlled by the given text.

We then ablate on the Top-K raito $\gamma_{\text{scope}}$ in Tab. 2: (1) within a proper range (experimentally, $\gamma_{\text{scope}} \in [0.2, 0.4]$) the results are quite robust; (2) the maximum normalization in Eq. 11 is important for the unity of high similarity and controllability, since different regions in the selected parts have different subject relevance and should be set to different weights. (3) Too small or too large influence scope will degrade similarity or controllability, respectively. These conclusions are validated by the visualization in Fig. 7.

**Effectiveness of *adaptive scoring module*.** As shown in Tab. 3, (1) We first compare with the simple use of all image features (ID-2), which results in degradation of both similarity and controllability, proving the importance of providing accurate and smooth influence quantity along with the coarse-to-fine diffusion generation process; (2) We then ablate on the module design (ID-3,4,5, ID-5), finding that using image score only results in worse performance. The reason is that the generation features are noisy at the beginning, resulting in an inaccurate score prediction. Therefore,

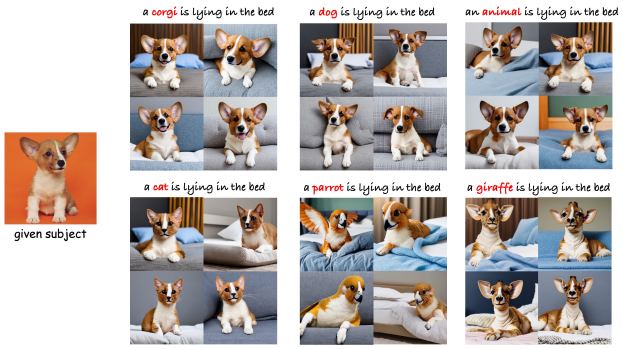


Figure 8. The customization by using different real text words.

we propose a step-scheduler to adaptively fuse text and image scores, leading to the best performance; (3) Finally, the choice of influence quantity $\gamma_{\text{num}}$ is ablated in ID-6 & 7.

**Impact of different read words.** The customization results in using different real text words are shown in Fig. 8. The real text word narrowed down for customization is highlighted in red. (1) No matter how coarse-grained text word is used, the customization results of *RealCustom* are **quite robust**. (2) When using *completely different word* to represent the given subject, *RealCustom* opens a door for a new application, *i.e.*, **novel concept creation**. That is, *RealCustom* will try to combine these two concepts and create a new one, *e.g.*, generating a parrot with the appearance of the given brown corgi, as shown in the below three rows. This application will be very valuable for designing new characters in movies or games, *etc*.

## 5. Conclusion

In this paper, we present a novel customization paradigm *RealCustom* that, for the first time, disentangles similarity of given subjects from controllability of given text by precisely limiting subject influence to relevant parts, which gradually narrowing the real word from its general connotation to the specific subject in a novel "train-inference" framework: the *adaptive scoring module* learns to adaptively modulate influence quantity during training; (2) the *adaptive mask guidance strategy* iteratively updates the influence scope and influence quantity of given subjects during inference. Extensive experiments demonstrate that RealCustom achieves the unity of high-quality similarity and controllability in the real-time open-domain scenario.

## 6. Acknowledgments

This work is supported by National Natural Science Foundation of China under Grant 62222212 and Science Fund for Creative Research Groups under Grant 62121002.
We would like to thank our colleagues at Bytedance, Wei Liu and Shiqi Sun, for their valuable help for this research.

## References


[1] Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. A neural space-time representation for text-to-image personalization. *arXiv preprint arXiv:2305.15391*, 2023. 2, 3

[2] 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. 3

[3] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. *arXiv preprint arXiv:2304.08465*, 2023. 3

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 6

[5] 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. 3

[6] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 1, 3

[7] Zhuowei Chen, Shancheng Fang, Wei Liu, Qian He, Mengqi Huang, Yongdong Zhang, and Zhendong Mao. Dreamidentity: Improved editability for efficient face-identity preserved image generation. *arXiv preprint arXiv:2307.00300*, 2023. 1, 3

[8] Giannis Daras and Alexandros G Dimakis. Multiresolution textual inversion. *arXiv preprint arXiv:2211.17115*, 2022. 2, 3

[9] Ziyi Dong, Pengxu Wei, and Liang Lin. Dreamartist: Towards controllable one-shot text-to-image generation via contrastive prompt-tuning. *arXiv preprint arXiv:2211.11337*, 2022.

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1, 2, 3, 6

[11] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Designing an encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2302.12228*, 2023. 2, 3

[12] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3

[13] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[14] 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. 3

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 6

[16] 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, 6

[17] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 2

[18] 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, 6

[19] Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. Divide & bind your attention for improved generative semantic nursing. *arXiv preprint arXiv:2307.10864*, 2023. 3

[20] 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

[21] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Guiding text-to-image diffusion model towards grounded generation. *arXiv preprint arXiv:2301.05221*, 2023. 3

[22] Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable image synthesis with multiple subjects. *arXiv preprint arXiv:2305.19327*, 2023. 2, 3

[23] 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. 4, 5

[24] 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, 3

[25] 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, 3

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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*, pages 234–241. Springer, 2015. 3

[27] 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, 2, 3, 6

[28] 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, 3

[29] 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. 6

[30] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 2, 3

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 6

[32] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. $p+$: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 2, 3

[33] Jinglong Wang, Xiawei Li, Jing Zhang, Qingyuan Xu, Qin Zhou, Qian Yu, Lu Sheng, and Dong Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *arXiv preprint arXiv:2309.02773*, 2023. 3

[34] 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, 6

[35] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. *arXiv preprint arXiv:2303.11681*, 2023. 3

[36] Changming Xiao, Qi Yang, Feng Zhou, and Changshui Zhang. From text to mask: Localizing entities using the attention of text-to-image diffusion models. *arXiv preprint arXiv:2309.04109*, 2023. 3

[37] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023. 6

[38] Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. Prospect: Expanded conditioning for the personalization of attribute-aware image generation. *arXiv preprint arXiv:2305.16225*, 2023. 2