EVENT-CUSTOMIZED IMAGE GENERATION

Anonymous authors
Paper under double-blind review

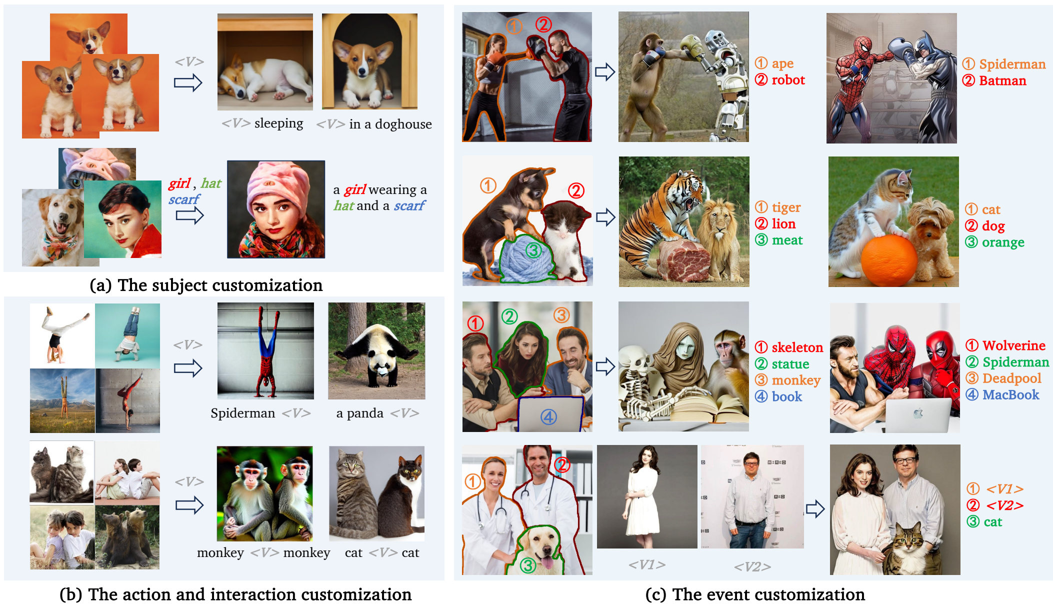

Figure 1: **Customized Image Generation.** **(a)** Generating customized images with given *subjects* in new contexts. **(b)** Generating customized images with co-existing basic *action* or *interaction* in given images. **(c)** Generating customized images for complex *events* with various target entities. Different colors and numbers show the associations between reference entities and their corresponding target prompts.

ABSTRACT

Customized Image Generation, generating customized images with user-specified concepts, has raised significant attention due to its creativity and novelty. With impressive progress achieved in *subject* customization, some pioneer works further explored the customization of *action* and *interaction* beyond entity (*i.e.*, human, animal, and object) appearance. However, these approaches only focus on basic actions and interactions between two entities, and their effects are limited by insufficient "exactly same" reference images. To extend customized image generation to more complex scenes for general real-world applications, we propose a new task: **event-customized image generation**. Given a single reference image, we define the "event" as all specific actions, poses, relations, or interactions between different entities in the scene. This task aims at accurately capturing the complex event and generating customized images with various target entities. To solve this task, we proposed a novel training-free event customization method: **FreeEvent**. Specifically, FreeEvent introduces two extra paths alongside the general diffusion denoising process: 1) Entity switching path: it applies cross-attention guidance and regulation for target entity generation. 2) Event transferring path: it injects the spatial feature and self-attention maps from the reference image to the target image for event generation. To further facilitate this new task, we collected two evaluation benchmarks: SWiG-Event and Real-Event. Extensive experiments and ablations have demonstrated the effectiveness of FreeEvent.

1 Introduction

Recently, large-scale pre-trained diffusion models (Dhariwal & Nichol, 2021; Nichol et al., 2021; Ramesh et al., 2022; Rombach et al., 2022; Saharia et al., 2022) have demonstrated remarkable success in generating diverse and photorealistic images from text prompts. Leveraging these unparalleled creative capabilities, a novel application — customized image generation (Gal et al., 2022; Ruiz et al., 2023; Chen et al., 2023) — has gained increasing attention for generating user-specified concepts. Significant progress has already been made in subject-customized image generation (Ye et al., 2023; Chen et al., 2024b). As shown in Figure 1(a), given a set of user-provided subject images, existing methods can accurately capture the unique appearance features of each subject (*e.g.*, `corgi`) with a special identifier token, enabling creative rendering in new and diverse scenarios. Moreover, they can seamlessly integrate multiple subjects into cohesive compositions, preserving their distinctive characteristics while adapting them to novel contexts.

Beyond the appearance of different entities (*i.e.*, humans, animals, and objects) in the images, pioneering approaches have been developed to customize the user-specified actions (Huang et al., 2024), interactive relations (Huang et al., 2023) and poses (Jia et al., 2024) between the entities. As shown in Figure 1(b), these methods attempt to capture the single-entity action (*e.g.*, `handstand`) or interactions (*e.g.*, `back to back`) between two entities that co-exist in the given reference images and transfer them to the synthesis of action- or interaction-specific images with new entities.

However, for real-world scenes that typically involve multiple entities with more complex interactions (*e.g.*, Figure 1(c), row three: `three humans are discussing in front of a computer with different poses`), these works (Huang et al., 2023; 2024; Jia et al., 2024) still face notable limitations. **1) Simplified Customization.** Current action customization (Huang et al., 2024) focuses solely on the basic actions of a single person. Similarly, interaction customizations (Huang et al., 2023; Jia et al., 2024) are limited to basic interactive relations or poses between just two entities. There is a lack of exploration into more complex and diverse actions or interactions that involve multiple humans, animals, and objects. Additionally, while these methods typically perform well when generating images with the same type of entity (*e.g.*, all monkeys or all cats), they struggle when faced with more diverse and complex entities and their combinations. These narrow focuses and limitation on entity generation have strictly limited their abilities to customize more complex and diverse scenes with creative content. **2) Insufficient Data.** To capture specific actions or interactions, existing methods (Huang et al., 2023; 2024; Jia et al., 2024) tend to represent them by learning corresponding identifier tokens, which can be further used for generating new images. However, for each action, or interaction, these training-based processes typically require a set of reference images (*e.g.*, 10 images) paired with corresponding textual descriptions across different entities. Unfortunately, each action or interaction is highly unique and distinctive, *i.e.*, gathering images that depict the exact same action or interaction is challenging. As shown in Figure 1(b), there are still significant differences in the same action (*e.g.*, `handstand`) between different reference images, which thus compromises the accuracy of learned tokens, leading to inconsistencies in action between generated images. This insufficient data issue for identical action or interaction has severely limited the practicality and generalizability of these methods.

To address these limitations and extend customized image generation to more complex scenes, we propose a new and meaningful task: **event-customized image generation**. Given a single reference image, we define the "event" as all actions and poses of each single entity, and their relations and interactions between different entities[1]. As shown in Figure 1(c), event customization aims to accurately capture the complex and diverse event from the reference image to generate target images with various combinations of target entities. Since it only needs one single reference image, the event customization also eliminates the need for collecting "exactly same" reference images.

To solve this challenging task, we proposed a novel *training-free* event customization method, denoted as **FreeEvent**. Based on the two main components of the reference image, *i.e.*, entity and event, FreeEvent decomposes the event customization into two parts: 1) Switching the entities in the reference image to target entities. 2) Transferring the event from the reference image to the target image. Following this idea, alongside the general denoising process of diffusion generation, we designed two extra paths: entity switching path and event transferring path. Specifically, entity switching path guides the localized layout of each target entity for entity generation. Event

[1] In this paper, we primarily measure the event complexity using the total number of entities.

transferring path further extracts the event information from the reference image and then injects it into the denoising process to generate the specific event. Through this direct guidance and injection, FreeEvent offers a significant advantage over existing methods by eliminating the need for time-consuming training. Furthermore, as shown in Figure 1(c), FreeEvent can also serve as a plug-and-play framework to combine with subject customization methods, generating creative images with both user-specified events and subjects.

Moreover, as a pioneering effort in this direction, we also collected two evaluation benchmarks from the existing dataset (*i.e.*, SWiG (Pratt et al., 2020) and HICO-DET (Chao et al., 2015)) and the internet for event-customized image generation, dubbed **SWiG-Event** and **Real-Event**, respectively. Both benchmarks include reference images featuring diverse events and entities, along with manually crafted target prompts. Extensive experiments demonstrate that our approach achieves state-of-the-art performance, enabling more complex and creative customization with enhanced practicality and generalizability.

In summary, we make several contributions in this paper: 1) We propose the novel event-customized image generation task, which extends customized image generation to more complex scenes in real-world applications. 2) We propose FreeEvent, the first training-free method for event customization, which can be further combined with subject customization methods for more creative and generalizable customizations. 3) We collect two evaluation benchmarks for event-customized image generation, and our FreeEvent achieves outstanding performance compared with existing methods.

2 Related Work

Text-to-Image Diffusion Generation. Diffusion models (Ho et al., 2020; Nichol & Dhariwal, 2021; Song et al., 2020) have emerged as a leading approach for image synthesis. The text-to-image diffusion models (Nichol et al., 2021; Ramesh et al., 2022; Saharia et al., 2022) further inject user-provided text descriptions into the diffusion process via pre-trained text encoders. After trained on large-scale text-image pairs, they have shown great success in text-to-image generation. Different from these models that operate the diffusion process on pixel space, the latent diffusion models (LDMs) (Rombach et al., 2022) propose to perform it on latent space with enhanced computational efficiency. Besides, existing works (Hertz et al., 2022; Tumanyan et al., 2023; Cao et al., 2023; Alaluf et al., 2024) have discovered the spatial feature and attention maps in LDMs contain localized semantic information of the image and the layout correspondence between textual conditions. As a result, these features and attention maps have been utilized to control the layout, structure, and appearance in text-to-image generation. This can be achieved either through a plug-and-play feature injection (Tumanyan et al., 2023; Xu et al., 2023; Lin et al., 2024) or by computing specific diffusion guidance (Epstein et al., 2023; Mo et al., 2024) for generation. In this paper, we utilize the pre-trained LDM Stable Diffusion (Rombach et al., 2022) as our base model.

Subject Customization. This task aims to generate customized images of user-specified subjects. Current mainstream subject customization works mainly focus on 1) Single subject customization, including learning specific identifier tokens (Gal et al., 2022), finetuning the text-to-image diffusion model (Ruiz et al., 2023; 2024), introducing layer-wise learnable embeddings (Voynov et al., 2023) and training large-scale multimodal encoders (Gal et al., 2023; Li et al., 2024). 2) Multi-subject composition, including cross-attention modification (Tewel et al., 2023), constrained model fine-tuning (Kumari et al., 2023), layout guidance (Liu et al., 2023), and gradient fusion of each subject (Gu et al., 2024). In conclusion, these works are all tailored to capture the appearance of the entities in the image, without considering the customization of actions or poses.

Action and Interaction Customization. They aim to generate customized images with co-existing actions or interactions in user-provided reference images. ReVersion (Huang et al., 2023) first proposes to customize specific interactive relations by optimizing the learnable relation tokens. ADI (Huang et al., 2024) makes progress in customizing specific actions for a single subject. And a following work (Jia et al., 2024) further extends it to learning interactive poses between two individuals. However, all these works only focus on simplified customization of some basic actions and interactions, and their effect is strictly limited by the insufficient data of reference images. In contrast, our proposed event customization only requires one reference image, and our training-free framework FreeEvent can achieve effective customization of complex events with various creative target entities. While the ImgAny (Lyu et al., 2024) also proposed a training-free framework for image generation through two branches, it focuses on the modeling of multi-modal inputs as conditions, which is beyond the scope of this paper.

3 METHODS

3.1 PRELIMINARY

Latent Diffusion Model. Generally, the LDMs include a pretrained autoencoder and a denoising network. Given an image x, the encoder $\mathcal{E}$ maps the image into the latent code $z_0 = \mathcal{E}(x)$, where the forward process is applied to sample Guassian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ to it to obtain $z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ from time step $t \sim [1, T]$ with a predefined noise schedule $\bar{\alpha}$. While the backward process iteratively removes the added noise on z_t to obtain z_0, and decodes it back to image with the decoder $x = \mathcal{D}(z_0)$. Specifically, the diffusion model is trained by predicting the added noise ϵ conditioned on time step t and possible conditions like text prompt P. The training objective is formulated as

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z\sim\mathcal{E}(x),P,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t; t, P)\|_2^2\right]. \tag{1}$$

where ϵ_θ is the denoising network.

Diffusion Guidance. The diffusion guidance modifies the sampling process (Ho et al., 2020) with additional score functions to guide it with more specific controls like object layout (Xie et al., 2023; Mo et al., 2024) and attributes (Epstein et al., 2023; Bansal et al., 2023). We express it as

$$\hat{\epsilon}_t = \epsilon_\theta(z_t; t, P) - s \cdot \mathbf{g}(z_t; t, P), \tag{2}$$

where $\mathbf{g}$ is the energy function and s is a parameter that controls the guidance strength.

3.2 TASK DEFINITION: EVENT-CUSTOMIZED IMAGE GENERATION

In this section, we first formally define the event-customized image generation task. Given a reference image I^R involves N reference entities $\mathrm{E}^R = \{R_1, \ldots, R_N\}$, we define the "event" as the specific actions and poses of each single reference entity, and the relations and interactions between different reference entities. Together we have the entity masks $\mathrm{M} = \{m_1, \ldots, m_N\}$, where m_i is the mask of its corresponding entity R_i. The event-customized image generation task aims to capture the reference event, and further generate a target image I^G under the same event but with diverse and novel target entities $\mathrm{E}^G = \{G_1, \ldots, G_N\}$ in the target prompt $\mathrm{P} = \{w_0, \ldots, w_N\}$, where w_i is the description of the target entity G_i, and each target entity G_i should keep the same action or pose with its corresponding reference entity R_i. As the example shown in Figure 2, given the reference image with four reference entities (*e.g.*, three people and one object), the event-customization aims to capture the complex reference event and generate the target image with a novel combination of different target entities (*e.g.*, skeleton, statue, monkey, book).

3.3 APPROACH

Overview. We now introduce the proposed training-free event customization framework FreeEvent. Specifically, we decompose the event-customized image generation into two parts, 1) generating target entities (*i.e.*, switching each reference entity to target entity), and 2) generating the same reference event (*i.e.*, transferring the event from the reference image to the target image). Following this idea, we design two extra paths for the diffusion denoising process of event customization, denoted as the entity switching path and the event transferring path, respectively. Generally, as shown in Figure 2, the generation of I^G starts by randomly initializing the latent $z_T^G \sim \mathcal{N}(0, \mathbf{I})$, and iteratively denoise it to z_0^G. During this denoising process, the entity switching path guides the generation of each target entity through cross-attention guidance and regulation based on the target prompt P and reference entity masks M. The event transferring path extracts the spatial features and self-attention maps from the reference image I^R, and then injects them to the denoising process. The final z_0^G is then transformed back to the target image I^G by the decoder.

U-Net Architecture The Stable Diffusion (Rombach et al., 2022) utilizes the U-Net architecture (Ronneberger et al., 2015) for ϵ_θ, which contains an encoder and a decoder, where each consists of several basic encoder/decoder blocks, and each encoder/decoder block further contains several encoder/decoder layers. Specifically, as shown in Figure 3(a), each U-Net encoder/decoder layer

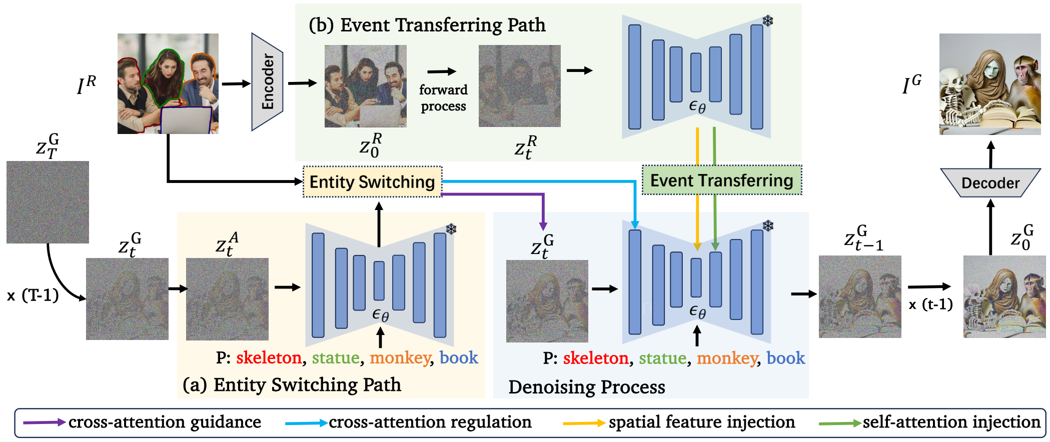

Figure 2: **The overview of pipeline.** Given the reference image, the event customization is overall a general diffusion denoising process with two extra paths. 1) The entity switching path guides the generation of each target entity through cross-attention guidance and regulation 2) The event transferring path injects the spatial features and self-attention maps from the reference image to the denoising process. The final z_0^G is then transformed back to target image I^G by the decoder.

consists of a residual module, a self-attention module, and a cross-attention module. For block b, layer l, and timestep t, the residual module produces the spatial feature of the image as $\mathbf{f}$. The self-attention module produces the self-attention map as $\mathbf{SA} = \mathrm{Softmax}(\frac{\mathbf{Q}_s\mathbf{K}_s^{\mathrm{T}}}{\sqrt{d}})$, where $\mathbf{Q}_s$ and $\mathbf{K}_s$ are query and key features projected from the visual features. For text-to-image generation, the cross-attention module further produces the cross-attention map between the text prompt P and the image as $\mathbf{CA} = \mathrm{Softmax}(\frac{\mathbf{Q}_c\mathbf{K}_c^{\mathrm{T}}}{\sqrt{d}})$, where $\mathbf{Q}_c$ is the query features projected from the visual features, and $\mathbf{K}_c$ is the key features projected from the textual embedding of P.

Entity Switching Path. This path aims on generating target entities $\mathrm{E}^G = \{G_1, \ldots, G_N\}$ in I^G by switching each refenrece entity R_i to G_i based on the target prompt P and reference entity masks M. And the key is to ensure each target entity G_i is generated at the same location as their corresponding reference entity R_i and avoid the appearance leakage between different entities. Inspired by prior works (Hertz et al., 2022; Chen et al., 2024a) that utilize the cross-attention maps to control the layout of text-to-image generation, we apply the cross-attention guidance and regulation to achieve the entity switching.

As shown in Figure 2(a), at the timestep t of the denoising process, we first obtain the latent for entity switching as $z_t^A = z_t^G$, we then input z_t^A together with the target prompt P into the U-Net, and calculate the cross-attention maps as $\mathbf{CA}^A$. As shown in Figure 3(b), we then introduce an energy function to bias the cross-attention of each token w_i as:

$$\mathbf{g}(\mathrm{CA}_i^A, m_i) = (1 - \frac{\mathrm{CA}_i^A * m_i}{\mathrm{CA}_i^A})^2 \tag{3}$$

where CA_i^A is the cross-attention map of token w_i. Optimizing this function encourages the cross-attention maps of each target entity G_i to obtain higher values inside the corresponding area specified by m_i, and further guide the localized layout of each target entity. We calculate the gradient of this guidance via backpropagation to update the latent z_t^G:

$$z_t^G = z_t^A - \sigma_t^2 \eta \bigtriangledown_{z_t^A} \sum_{i \in N} \mathbf{g}(\mathrm{CA}_i^A, m_i) \tag{4}$$

where η is the guidance scale and $\sigma_t = \sqrt{(1 - \bar{\alpha}_t / \bar{\alpha}_t}$. Additionally, to avoid the appearance leakage between each target entity, we further regulate the cross-attention map of each token within its corresponding area. Specifically, for cross-attention maps $\mathbf{CA}^G$ calculated at timestep t during the denoising process, we have:

$$\mathrm{CA}_i^G = m_i \odot \mathrm{CA}_i^G \tag{5}$$

where CA_i^G is the cross-attention map of token w_i.

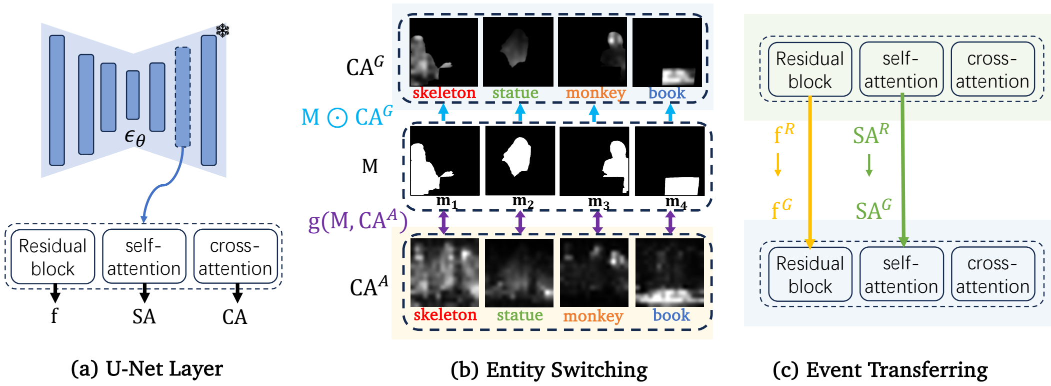

Figure 3: **(a)** The architecture of the U-Net layer. **(b)** The process of cross-attention guidance and regulation. **(c)** The process of spatial feature and self-attention injection.

Event Transferring Path. This path aims to extract the specific reference event from the reference image I^R, including the action, pose, relation, or interactions between each reference entity, and transferring them to the target image I^G. Meanwhile, from the perspective of image spatial information, the event is essentially the structural, semantic layout, and shape details of the image. Thus, based on the observation that the spatial features and self-attention maps can be utilized to control the image layout and structure (Tumanyan et al., 2023; Xu et al., 2023; Lin et al., 2024), we perform spatial feature and self-attention map injection to achieve the event transferring.

Specifically, as shown in Figure 2(b) we first get the latent code of the reference image $z_0^R = \mathcal{E}(I^R)$, and at each time step t during the denoising process, we obtain z_t^R via the diffusion forward process. We then input z_t^R into the U-Net to extract the spatial features and self-attention maps of the reference image as $\mathbf{f}^R$ and $\mathbf{SA}^R$. Parallelly, for the denoising process, we input z_t^G together with the target prompt P into the U-Net, and calculate the spatial features and self-attention maps for the generated image as: $\mathbf{f}^G$ and $\mathbf{SA}^G$. Then, as shown in Figure 3(c), we perform the injection by directly replacing corresponding spatial features and self-attention maps:

$$\mathbf{f}^G \leftarrow \mathbf{f}^R \quad \text{and} \quad \mathbf{SA}^G \leftarrow \mathbf{SA}^R. \tag{6}$$

Highlights. By applying cross-attention guidance and regulation on each text token, our attention-guided entity switching can also be used to generate target entities of user-specified subjects, *i.e.*, represented by specific identifier tokens. Thus, our framework can be easily combined with subject customization methods to generate creative images with both customized events and subjects.

4 Experiments

4.1 Experimental Setup

Evaluation Benchmarks. In order to provide sufficient and suitable conditions for both quantitative and qualitative comparisons on this new task, we collect two new benchmarks[2]. 1) For quantitative evaluation, we present *SWiG-Event*, a benchmark derived from SWiG (Pratt et al., 2020) dataset, which comprises 5,000 samples with various events and entities, *i.e.*, 50 kinds of different actions, poses, and interactions, where each kind of event has 100 reference images, and each reference image contains 1 to 4 entities with labeled bounding boxes and nouns. 2) For qualitative evaluation, we present *Real-Event*, which comprises 30 high-quality reference images from HICO-DET (Chao et al., 2015) and the internet with a wide range of events and entities (*e.g.*, animal, human, object, and their combinations). We further employ Grounded-SAM (Kirillov et al., 2023; Ren et al., 2024) to extract the mask of each entity.

Baselines. To evaluate the effectiveness of our method, we compared it with several kinds of state-of-the-art baselines. For conditioned text-to-image generation baselines, we compared with the training-based method ControlNet (Zhang et al., 2023) and the training-free method BoxDiff (Xie

[2]Due to the limited space, more details are left in the Appendix.

Model	Image Retrieval			Verb Detection			CLIP Score		FID↓
	R@1↑	R@5↑	R@10↑	T-1↑	T-5↑	T-10↑	CLIP-I↑	CLIP-T↑	
ControlNet	10.64	26.12	36.82	10.66	23.98	31.28	0.6009	0.2198	70.45
BoxDiff	8.60	22.48	32.08	5.58	14.52	19.42	0.5838	0.2153	68.49
FreeEvent	**41.12**	**63.02**	**72.74**	**34.10**	**62.04**	**71.82**	**0.7044**	**0.2238**	**29.05**

Table 1: Performance of our model and state-of-art conditional text-to-image generation models on SWiG-Event. For image retrieval, the R@k represents that among the top-k images with the highest similarity to the target image, its corresponding reference image is included. For verb detection, the T-K represents the top-k detection accuracy.

et al., 2023). For localized editing baselines, we compared with training-free methods PnP (Tumanyan et al., 2023) and MAG-Edit (Mao et al., 2023). For customization baselines, we compared with training-based methods Dreambooth (Ruiz et al., 2023) and ReVersion (Huang et al., 2023).

Implementation Details. We use Stable Diffusion v2-1-base as the base model for all methods, and images are generated with a resolution of 512x512 on a NVIDIA A100 GPU[2].

4.2 Quantitative Comparisons

In this subsection, we compare our method with conditional text-to-image generation baselines ControlNet (Zhang et al., 2023) and BoxDiff (Xie et al., 2023) on the SWiG-Event benchmark.

Setting. Each reference image in SWiG-Event contains reference entities together with labeled event class, bounding boxes, nouns, and their corresponding masks. Specifically, we construct the target prompt as a list of reference entity nouns, *i.e*., we ask all the methods to *reproduce* the event of the reference image with the same reference event and same reference entities. Additionally, ControlNet takes the semantic map merged from the masks as the layout condition, and BoxDiff takes the bounding boxes with labeled entity nouns as the layout condition[2].

Evaluation. We apply multiple metrics to evaluate the customization quality of 5,000 target images. 1) Image retrieval performance. We retrieved each target image for its corresponding reference image based on the CLIP score across all the 100 reference images that have the same reference event class. Specifically, we extracted the image feature of each image through a pre-trained CLIP (Radford et al., 2021) visual encoder and calculated the cosine similarities for image retrieval. 2) Verb detection performance. We utilized the verb detection model GSRTR (Cho et al., 2021) which was trained on the SWIG dataset to detect the verb class of each generated image, and then calculated the detection accuracy based on the annotations of the reference images (*i.e*., whether the generated images and their reference images have the same verb class). 3) Standard image generation metrics. For a more comprehensive comparison, we used the FID (Heusel et al., 2017) score, the CLIP-I (Radford et al., 2021) score, and the CLIP-T (Radford et al., 2021) score. We use the CLIP-I score to evaluate the image alignment of generated images with their reference images. And use the CLIP-T score to evaluate the text alignment of the generated images with text prompts.

Results. As shown in Table 1, we can observe: 1) FreeEvent has better retrieval performance than both ControlNet and BoxDiff. This demonstrates that the target images generated by FreeEvent better preserve the overall characteristics of the reference event and entity. 2) FreeEvent also achieves the best verb detection performance, which indicates our method can better preserve the interaction semantics of the generated images. 3) FreeEvent further achieves superior performance over baselines across all standard image generation metrics, indicating our method can generate images with better qualities and alignment with both the reference images and texts. These results all demonstrate the effectiveness of FreeEvent for event customization.

4.3 Qualitative Comparisons

We compare FreeEvent with a wide range of state-of-the-art baselines on the Real-Event benchmark, including conditioned text-to-image generation method **ControlNet** (Zhang et al., 2023) and **BoxDiff** (Xie et al., 2023), localized image editing method **PnP** (Tumanyan et al., 2023) and **MAG-Edit** (Mao et al., 2023), image customization methods **Dreambooth** (Ruiz et al., 2023) and **ReVersion** (Huang et al., 2023).

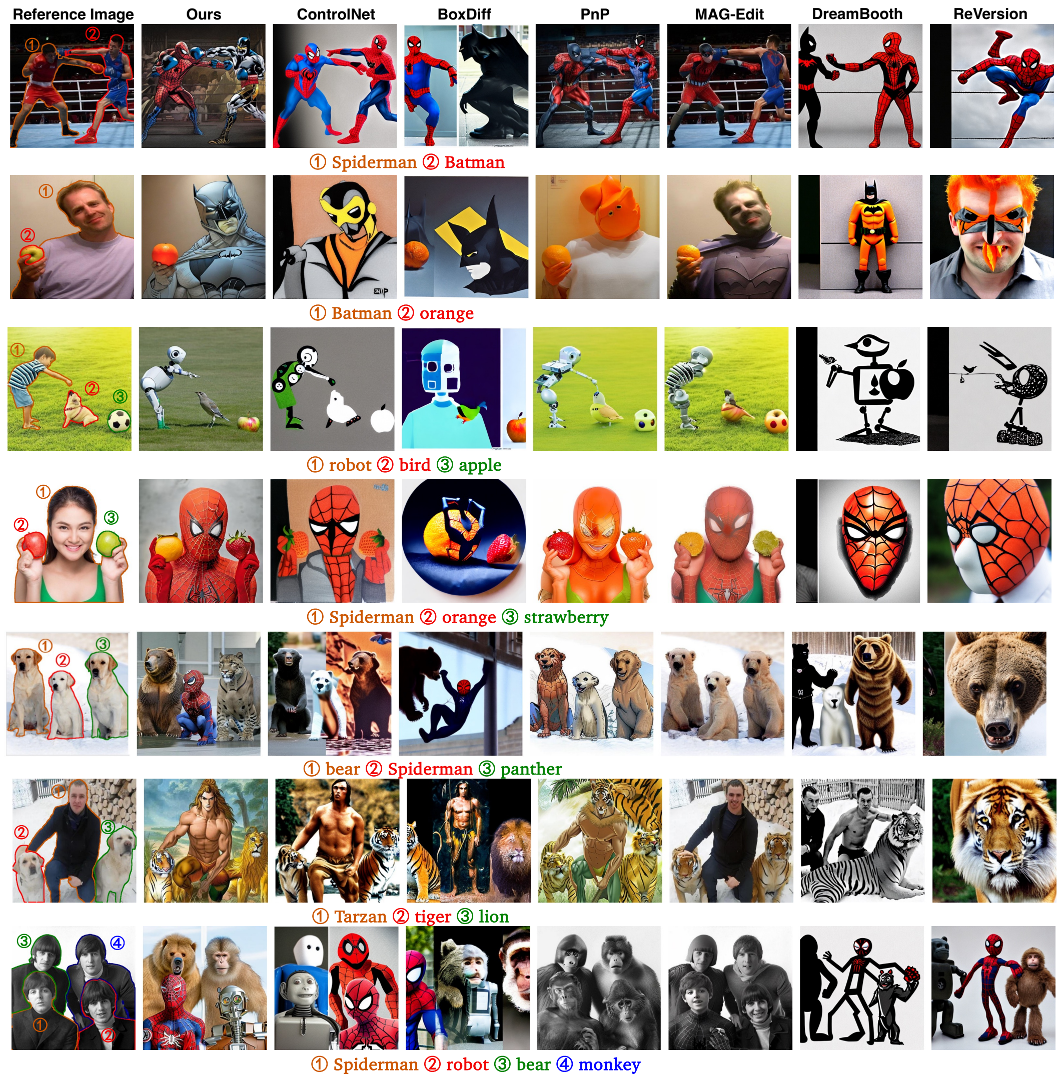

Figure 4: **Comparision of Event Customization.** Different colors and numbers show the associations between reference entities and their corresponding target prompts.

Setting. For each reference image in Real-Event, we manually constructed target prompts with various combinations of different target entities. Specifically, ControlNet takes the semantic map and BoxDiff takes the labeled bounding boxes as the layout conditions. MAG-Edit takes the reference entity masks for localized editing. Dreambooth and ReVersion learn event-specific identifier tokens for text-to-image generation.

Results. As shown in Figure 4, we can observe: 1) Conditional text-to-image generation models ControlNet and BoxDiff can only maintain the rough layout of each entity and struggle to capture the detailed action, pose, or interaction between different entities. And they both failed to match the generated entity with the desired target prompt. 2) For localized image editing methods PnP and MAG-Edit, while they can capture the reference event, they both struggle to accurately generate the target entities, and suffer from severe appearance leakage between each target entity (*e.g.*, `orange` and `strawberry` in row four, `tiger` and `lion` in row six), and sometimes even failed to edit and output the original content. 3) The subject-customization model Dreambooth and the relation-customization model ReVersion both failed to generate satisfying results. As discussed before, these training-based methods require multiple reference images and are unable to learn the specific event when facing only one reference image. 4) Obviously, our FreeEvent successfully achieves the customization of various complex events with novel combinations of target entities. Meanwhile, the ControlNet and the localized image editing models tend to generate the target entities strictly

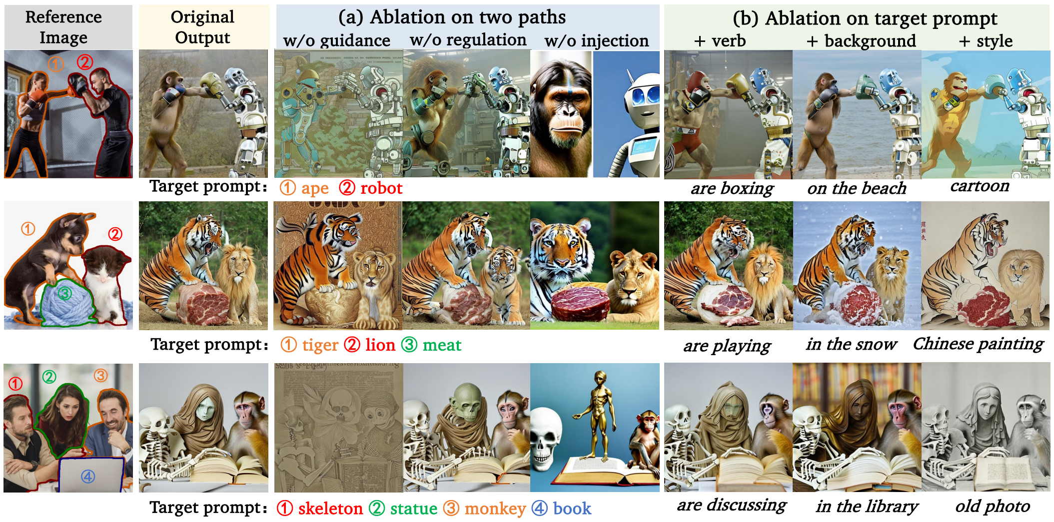

Figure 5: **Ablations of the proposed paths and the target prompt.** The “guidance” and “regulation” denote the cross-attention guidance and cross-attention regulation in the entity switching path, respectively. The “injection” denotes the event transferring path.

matching the mask of their corresponding reference entities (*e.g.*, `bird` in row three), which appears very incongruous. On the contrary, the entities generated by FreeEvent not only match the layout of the reference entity but also keep it harmonious. After all, while we use the reference entity mask to guide the generation of each target entity, the cross-attention guidance focuses on directing the overall layout of each target entity and does not restrict their detailed appearance, allowing for a more diverse generation of target entities[2].

4.4 Ablations

Effectiveness of Entity Switching Path and Event Transferring Path. We first run ablations to verify the effect of two proposed paths during event customization.

Results. As the results are shown in Figure 5(a), we can observe: 1) For the entity switching path, removing the cross-attention guidance results in the failure of target entities generation (*e.g.*, the ape, the meat), and removing cross-attention regulation leads to the appearance leakage between entities (*e.g.*, the tiger and lion, the skeleton and statue). 2) After removing the event transferring path, although the target entities can be generated, the reference events are completely lost (*i.e.*, the pose, action, relations, and interactions between each entity). These results all corroborates the effect of two paths in event customization.

Influence of Different Target Prompts. Noteablly, in our paper, the target prompt only contains the nouns of the target entities, we then run the ablations to analyze the influence of different descriptions (*i.e.*, verb, background, style) in the target prompt for event customization.

Results. From Figure 5(b) we can observe: 1) Adding *verb* description leads to a certain degree of negative impact on entity appearance (*e.g.*, the head of the ape, the face of the monkey) since these verbs may not be aligned with the model. Besides, accurately describing events in complex scenes can be challenging for users. Therefore, since FreeEvent can already achieve precise extraction and transfer of the reference events, users do not need to describe the specific events in the target prompt, which further demonstrates FreeEvent’s practicality. 2) FreeEvent can accurately generate extra contents for the *background* and *style*. Although there may be some detailed changes in the entity’s appearance compared to the original output, these do not affect the entity’s characteristics or the event. This also demonstrates FreeEvent’s strong generalization capability.

Combination of Event and Subject Customization. We further validate the ability of our framework to combine with subject customization methods to generate target entities with user-specified subjects, *i.e.*, represented by identifier tokens. We took the Break-A-Scene model (Avrahami et al., 2023) to learn identifier tokens for each subject and replaced the Stable Diffusion models in Figure 2 with the fine-tuned one.

Figure 6: **Results of Event-Subject Customization.** Different colors and numbers show the associations between reference entities and their corresponding target prompts.

Model	Ours	ControlNet	BoxDiff	PnP	MAG-Edit	DreamBooth	ReVersion
HJ	**48**	19	2	31	13	1	0

Table 2: Results of the user studies on the Real-Event. "HJ" denotes the count of human judgment.

Results. As shown in Figure 6, FreeEvent can effectively generate various given subjects in specific events. Specifically, FreeEvent enables the flexible generation of a wide range of subject concepts (*e.g.*, humans, regular objects, and backgrounds) and their combinations. These results demonstrated the great potential of our framework for Event-Subject customization.

4.5 User Study

Setting. We conducted user studies on Real-Event to further evaluate the effectiveness of FreeEvent. Specifically, we invited 10 experts and gave them a reference image, a target prompt, and seven target images generated by different models. They are asked to choose the three target images that they believe demonstrate the best results in event customization, taking into account the generation effects of the events and entities, as well as the overall coherence of the images. We prepared 50 trials and asked the experts to give their judgments. The target image which got more than six votes is regarded as human judgment.

Results. As shown in Table 2, FreeEvent achieves better performance on human judgments (HJ) compared with all the baseline models.

5 Conclusion

In this paper, we proposed a new image generation task: Event-Customized Image Generation. It focuses on the customization of complex events with various target entities. Meanwhile, we proposed the first training-free event-customization framework **FreeEvent**. To facilitate this new task, we also collected two evaluation benchmarks from existing datasets and the internet, dubbed SWiG-Event and Real-Event, respectively. We validate the effectiveness of FreeEvent with extensive comparative and ablative experiments. Moving forward, we are going to 1) extend the event customization into other modalities, *e.g.*, video generation; 2) explore advanced techniques for the finer combination of different customization works, *e.g.*, subject, event, and style customizations.

REFERENCES

Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024.

Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–12, 2023.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 843–852, 2023.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22560–22570, 2023.

Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE international conference on computer vision*, pp. 1017–1025, 2015.

Hong Chen, Yipeng Zhang, Simin Wu, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. *arXiv preprint arXiv:2305.03374*, 2023.

Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5343–5353, 2024a.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6593–6602, 2024b.

Junhyeong Cho, Youngseok Yoon, Hyeonjun Lee, and Suha Kwak. Grounded situation recognition with transformers. In *British Machine Vision Conference (BMVC)*, 2021.

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

Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023.

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.

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

Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

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.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

Siteng Huang, Biao Gong, Yutong Feng, Xi Chen, Yuqian Fu, Yu Liu, and Donglin Wang. Learning disentangled identifiers for action-customized text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7797–7806, 2024.

Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu. Reversion: Diffusion-based relation inversion from images. *arXiv preprint arXiv:2303.13495*, 2023.

Xu Jia, Takashi Isobe, Xiaomin Li, Qinghe Wang, Jing Mu, Dong Zhou, Huchuan Lu, Lu Tian, Ashish Sirasao, Emad Barsoum, et al. Customizing text-to-image generation with inverted interaction. In *ACM Multimedia 2024*, 2024.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

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*, pp. 1931–1941, 2023.

Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024.

Kuan Heng Lin, Sicheng Mo, Ben Klingher, Fangzhou Mu, and Bolei Zhou. Ctrl-x: Controlling structure and appearance for text-to-image generation without guidance. *arXiv preprint arXiv:2406.07540*, 2024.

Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. *arXiv preprint arXiv:2303.05125*, 2023.

Yuanhuiyi Lyu, Xu Zheng, and Lin Wang. Image anything: Towards reasoning-coherent and training-free multi-modal image generation. *arXiv preprint arXiv:2401.17664*, 2024.

Qi Mao, Lan Chen, Yuchao Gu, Zhen Fang, and Mike Zheng Shou. Mag-edit: Localized image editing in complex scenarios via mask-based attention-adjusted guidance. *arXiv preprint arXiv:2312.11396*, 2023.

Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7465–7475, 2024.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pp. 314–332. Springer, 2020.

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*, pp. 8748–8763. PMLR, 2021.

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.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

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*, pp. 10684–10695, 2022.

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*, pp. 234–241. Springer, 2015.

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*, pp. 22500–22510, 2023.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6527–6536, 2024.

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.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–11, 2023.

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

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.

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7452–7461, 2023.

Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with natural language. *arXiv preprint arXiv:2312.04965*, 2023.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

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

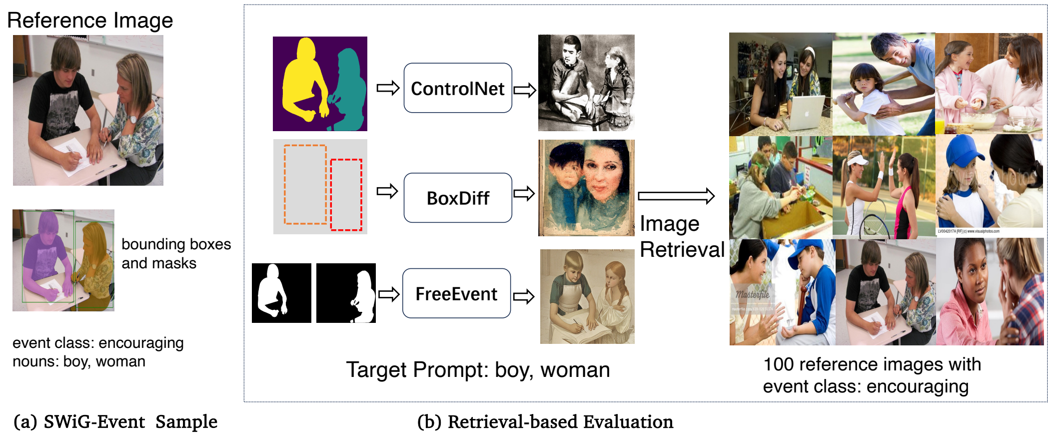

Figure 7: **(a)** The SWiG-Event sample. **(b)** The process of quantitative evaluation and image retrieval.

APPENDIX

The Appendix is organized as follows:

- In Sec. A, we show more implementation details.
- In Sec. B, we show more details of the SWiG-Event benchmark and the process of quantitative evaluation and image retrieval.
- In Sec. C, we show the results for attribute generation during event customization. .
- In Sec. D, we provide the discussion of our work's limitations and potential negative societal impacts.
- In Sec. E, we show more qualitative comparison results of event customization on the Real-Event. .

A IMPLEMENTATION DETAILS.

The denoising process was set with 50 steps. For entity switching path, for all blocks and layers containing the cross-attention module, we apply the cross-attention guidance during the first 10 steps. And apply the cross-attention regulation during the whole 50 steps. For event transferring path, we perform spatial feature injection for block and layer at $\{\text{decoder block } 1: [\text{layer } 1]\}$ during the whole 50 steps. And perform self-attention injection for blocks and layers at $\{\text{decoder block } 1: [\text{layer } 1, 2], \text{decoder block } 2: [\text{layer } 0, 1, 2], \text{decoder block } 3: [\text{layer } 0, 1, 2]\}$ during the first 25 steps. We set the classifier-free guidance scale to 15.0.

B DETAILS OF SWIG-EVENT AND PROCESS OF IMAGE RETRIEVAL.

As shown in Figure 7(a), each SWiG-Event sample consists of a reference image with labeled bounding boxes and masks for each reference entity, the nouns of each reference entity, and the event class. As shown in Figure 7(b), we constructed the target prompt as a list of reference entity nouns. The ControlNet takes the semantic map merged from the masks as the layout condition, and BoxDiff takes the bounding boxes with labeled entity nouns as the layout condition.

To compare the image retrieval performance, we retrieved the target image for its corresponding reference image across all the 100 reference images that have the same reference event class.

C ATTRIBUTE GENERATION RESULTS.

In this paper, we didn't explicitly model the attributes during generation. However, as the results are shown in Figure 5(b), since we can generate extra content for background and style by giving

Figure 8: The results of attribute generation during event customization.

corresponding text descriptions, we thus tried to model the attributes by giving extra adjectives to the target prompt as an easy and natural exploration. Meanwhile, to ensure the accurate generation of the attributes, we applied the cross-attention guidance and regulation on each attribute using the mask of the entity they describe. As the results shown in Figure 8, our method successfully addresses the attributes of the corresponding entity (*e.g.*, colors, materials, and ages). After all, while the attribute part is not the primary focus of our work, our approach shows potential and effectiveness in addressing it, and we would be happy to conduct further research in our future work.

D Limitation and Potential Negative Societal Impact.

Limitations. The main limitation of FreeEvent lies in the complexity of events and the number of entities. The customization effect may be compromised when there are too many entities in an image, especially if they are too small. As the first work in this direction, we hope our method can unveil new possibilities for more complex customization and the generation of a greater number of richer, more diverse entities. Additionally, since our model is built on pretrained Stable Diffusion (SD) models, our performance depends on the generative capabilities of SD. This can lead to suboptimal results for entities that the current SD struggles with, such as human faces and hands.

Potential Negative Societal Impacts. Since FreeEvent can seamlessly integrate with subject customization methods to generate target entities based on user-specified subjects, this capability also raises the same concerns about the potential misuse of pretrained SD models for malicious applications (*e.g.*, Deepfakes) involving real human figures. To address this, it is essential to implement robust safeguards and ethical guidelines, similar to the security measures and NSFW content detection mechanisms already present in existing diffusion models.

E More Qualitative Comparision Results.

We show more comparisons on Real-Event in Figure 9, Figure 10, Figure 11, Figure 12 and Figure 13. Specifically, we list them by the order of entity numbers. And we use different combinations of target entities for the same reference image to generate diverse target images.

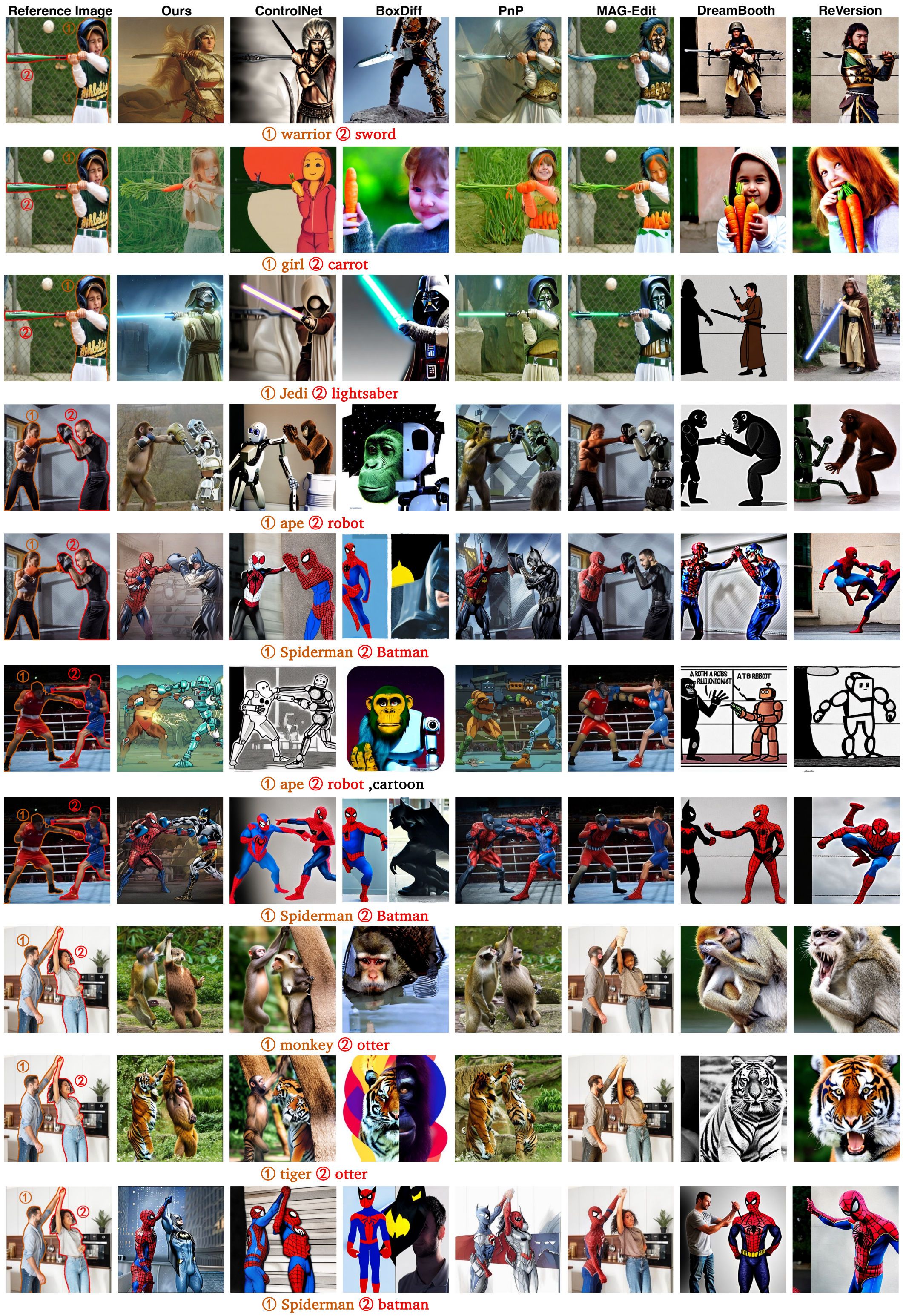

Figure 9: **Comparision of Event Customization.** Different colors and numbers show the associations between reference entities and their corresponding target prompts.

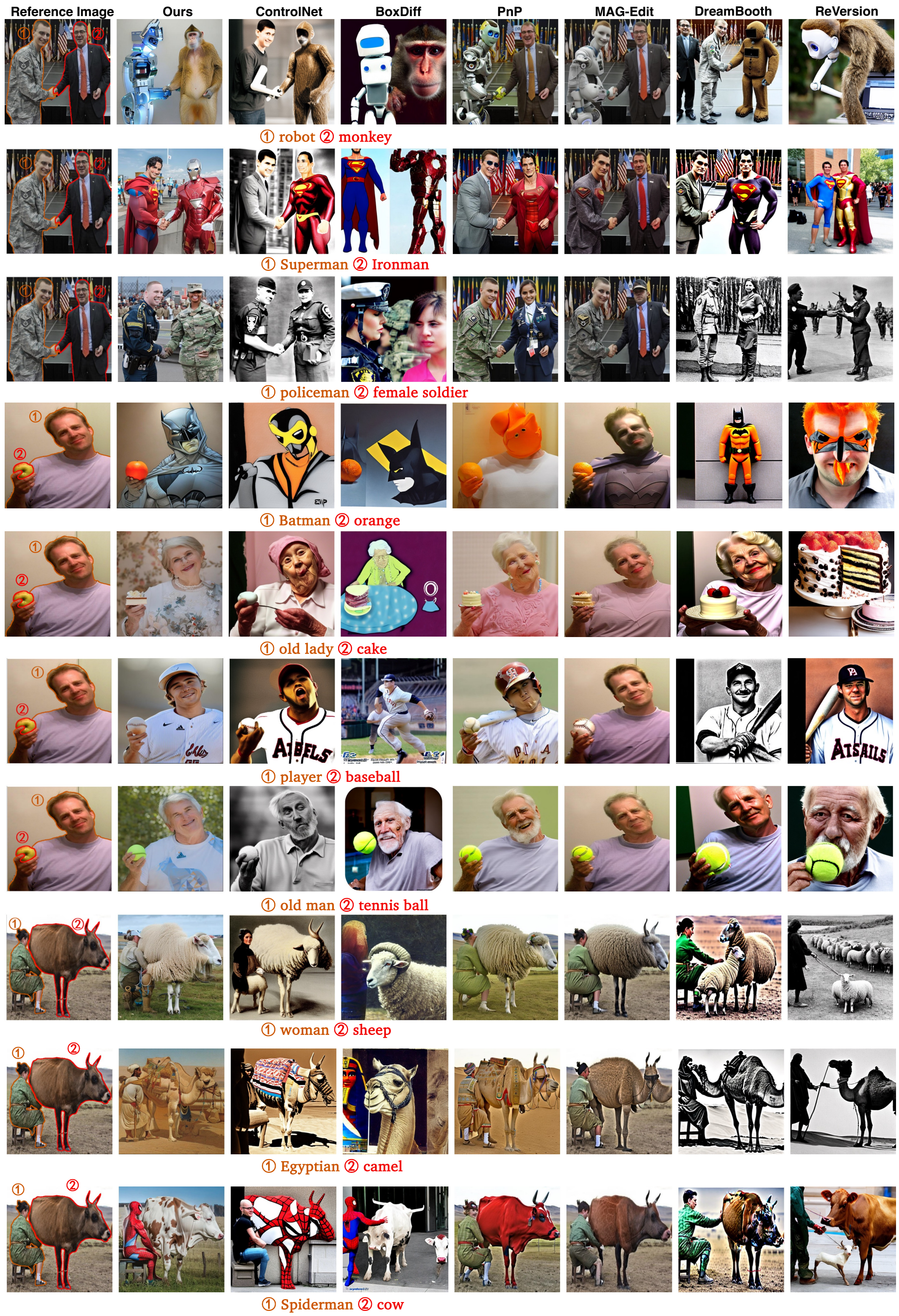

Figure 10: **Comparision of Event Customization.** Different colors and numbers show the associations between reference entities and their corresponding target prompts.

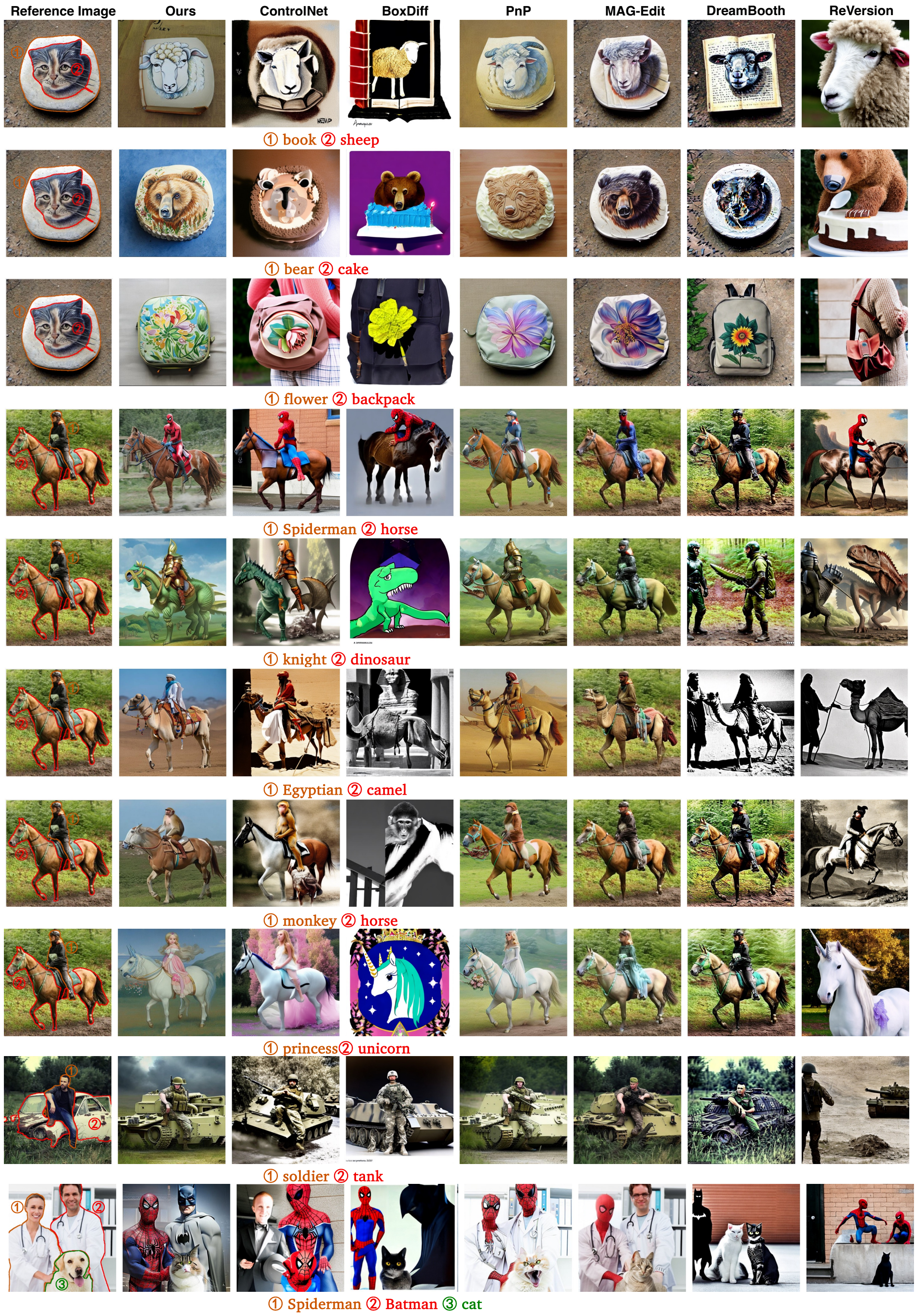

Figure 11: **Comparision of Event Customization.** Different colors and numbers show the associations between reference entities and their corresponding target prompts.

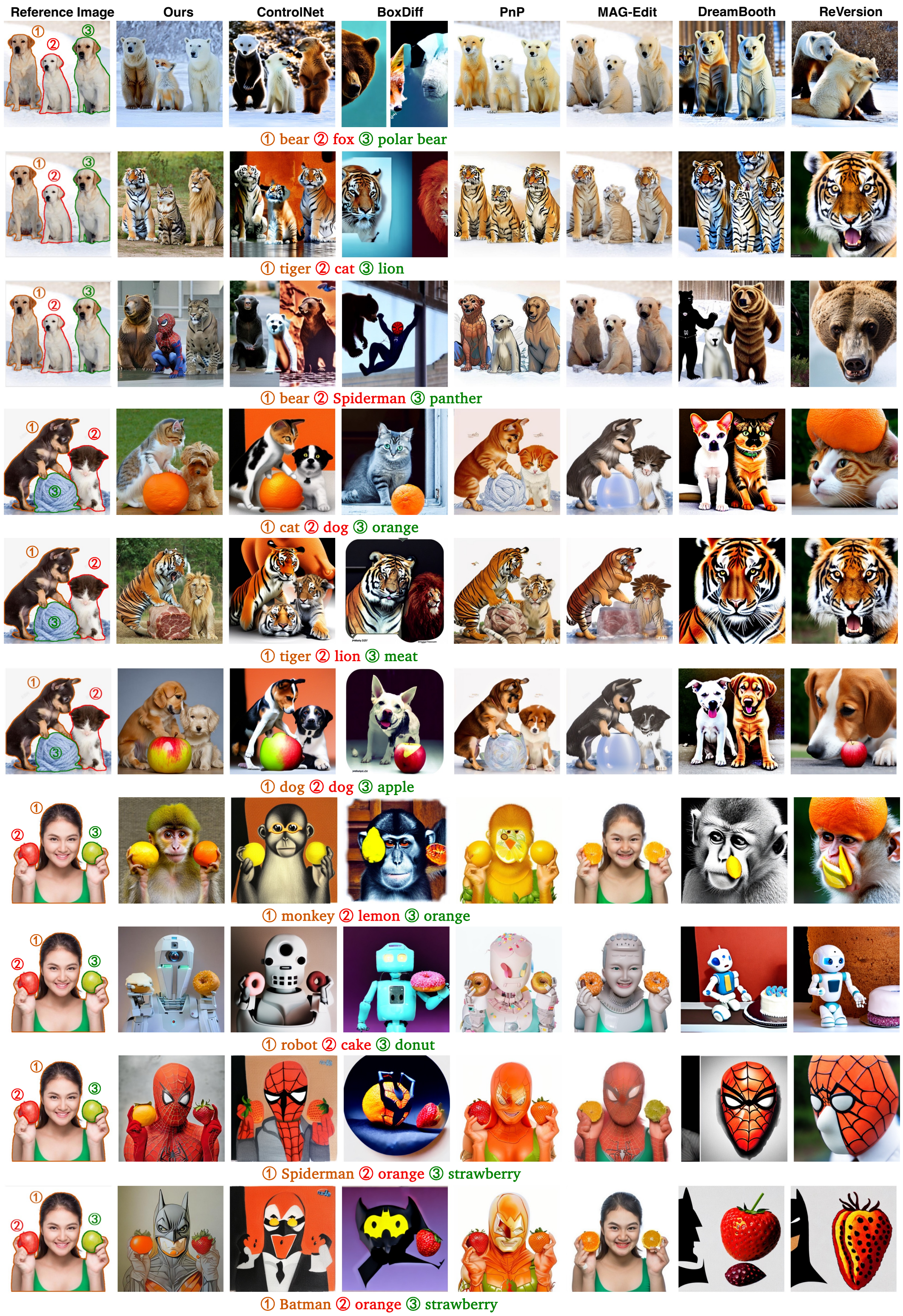

Figure 12: **Comparision of Event Customization.** Different colors and numbers show the associations between reference entities and their corresponding target prompts.

Figure 13: **Comparision of Event Customization.** Different colors and numbers show the associations between reference entities and their corresponding target prompts.