Harmonizing Visual and Textual Embeddings for Zero-Shot Text-to-Image Customization

Yeji Song, Jimyeong Kim*, Wonhark Park*, Wonsik Shin, Wonjong Rhee, Nojun Kwak†

Seoul National University
{ldynx, wlaud1001, pwh0515, wonsikshin, wrhee, nojunk}@snu.ac.kr

Abstract

In a surge of text-to-image (T2I) models and their customization methods that generate new images of a user-provided subject, current works focus on alleviating the costs incurred by a lengthy per-subject optimization. These zero-shot customization methods encode the image of a specified subject into a visual embedding which is then utilized alongside the textual embedding for diffusion guidance. The visual embedding incorporates intrinsic information about the subject, while the textual embedding provides a new context. However, the existing methods often 1) generate images with the same pose as an input image, and 2) exhibit deterioration in the subject's identity when facing a pose variation prompt. We first pin down the problem and show that redundant pose information in the visual embedding interferes with the pose indication in the textual embedding. Conversely, the textual embedding also harms the subject's identity which is tightly entangled with the pose in the visual embedding. As a remedy, we propose *text-orthogonal visual embedding* which effectively harmonizes with the given textual embedding. We also adopt the visual-only embedding and inject the subject's clear features using a *self-attention swap*. Our method is both effective and robust, offering highly flexible zero-shot generation while effectively maintaining the subject's identity.

Extended version — https://arxiv.org/abs/2403.14155

Introduction

Recent advancements in text-to-image (T2I) generation, especially diffusion models (Rombach et al. 2022; Balaji et al. 2023; Saharia et al. 2022; Ramesh et al. 2022; Nichol et al. 2022) have opened up a new era of image creation. Subject-driven generation (Ruiz et al. 2023; Gal et al. 2022a; Tewel et al. 2023a; Qiu et al. 2023; Voynov et al. 2023) aims to generate novel images featuring a specific subject provided by the user. The common approach represents the subject as a new pseudo-word (S^*) in the textual embedding space of the text encoder. They optimize a pseudo-word by updating the textual embedding of the pseudo-word (Gal et al. 2022b) or the diffusion model's parameters (Ruiz et al. 2023; Tewel et al. 2023b). However, these approaches often do not align with the actual needs of users, who typically 1) possess constrained GPU resources, 2) desire real-time applications, and 3) pursue a convenient rendition with a single input image. In response to these challenges, single-image-based *zero-shot* customization methods (Li, Li, and Hoi 2024; Wei et al. 2023; Chen et al. 2023; Ma et al. 2023; Jia et al. 2023; Yuan et al. 2023; Xiao et al. 2023; Zhang et al. 2024; Song et al. 2025; Ye et al. 2023) have been proposed.

To obtain the pseudo-word from a single image without the time and resource-intensive optimization process, they adopt the pre-trained mappers such as MLP network (Wei et al. 2023; Yuan et al. 2023; Xiao et al. 2023), adapter (Chen et al. 2023; Shi et al. 2023; Ma et al. 2023) or multi-modal encoder (Li, Li, and Hoi 2024) to transform a subject's image into the visual embedding. The visual embedding provides representative information about the subject's identity from the input image, utilized along with the textual embedding which contains a novel desired context. This approach eliminates the need for additional training processes while effectively depicting the subject in new scenes.

However, the existing methods are susceptible to confusing the subject's identity with other irrelevant details within an image. To remedy this, they employ the subject's segmentation mask (Li, Li, and Hoi 2024; Wei et al. 2023; Ma et al. 2023; Jia et al. 2023; Xiao et al. 2023) or utilize the embedding from the deepest layers of an image encoder (Wei et al. 2023), which effectively separate the subject from the background or other surrounding objects. However, they cannot disentangle the subject's pose from its identity as these aspects are tightly intertwined within the same pixels. As shown in Figure 1, when attempting to change the subject's pose, the generated subject either remains in the same pose as the input image or partially loses its identity. We term this phenomenon as the *pose-identity entanglement* within the visual embedding, which bottlenecks the diverse applications of customization methods.

Specifically, the pose-identity entanglement prompts the visual embedding to carry the visual features of the subject *in a specific pose*. When using a pose-related text prompt, the input image and the text prompt provide two concurrent but conflicting embeddings to the model with different pose information. Therefore, a conflict occurs between the visual and textual embeddings, impeding their functions. To verify this conflict, we conducted the experiment with BLIP-

*These authors contributed equally.
†Corresponding Author

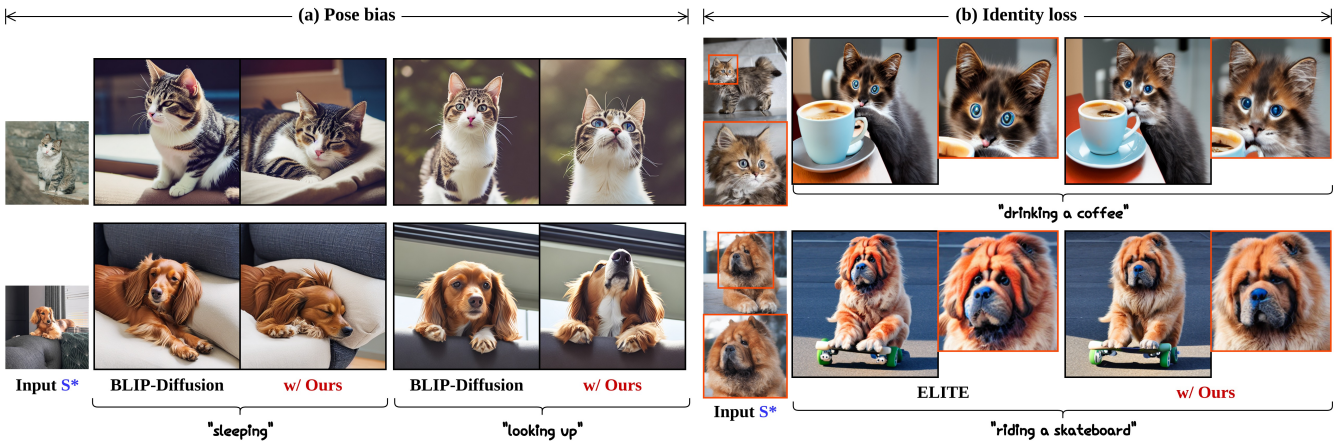

Figure 1: The baselines struggle with either (a) strong bias toward the pose in the input image, (b) loss in the subject's identity, or both. Our method deals with these challenges and paves the way for a more diverse and lively subject-driven generation.

Diffusion (Li, Li, and Hoi 2024) to generate images using only the textual embedding while zero-padding the image embedding. We found that this results in the subject with various poses faithfully following the text prompt, leading to a better CLIP-T score (+0.033). Using only the visual embedding also results in the images retaining the subject's whole identity, leading to a better CLIP-I score (+0.047).

Regarding this, we have identified two significant problems stemming from the conflict between the visual and textual embeddings when modifying the subject's pose:

- **Pose bias:** the generated images tend to maintain the original pose of the subject presented in the input image.
- **Identity loss:** the subject in the generated images partially loses its identity, appearing with a different color or body shape.

The visual embedding readily interferes with the textual embedding, causing the *pose bias*. Conversely, the textual embedding also interferes with visual embedding, resulting in *identity loss*. In this paper, we focus on resolving these two problems. They are imperative tasks for advancing towards a more diverse and desirable customization, while highly challenging due to pose-identity entanglement.

To alleviate this conflict, we propose *contextual embedding orchestration* and *self-attention swap*. The former involves adjusting the visual embedding to align better with the textual embedding by orthogonalizing it to the subspace of the textual embedding vectors. The latter takes advantage of another denoising process guided by the visual-only embedding that fundamentally evades the conflict and aggregates the subject's clean information. Our method is generic and easily applicable to any zero-shot customization method that utilizes visual and textual embeddings, since it does not require an additional tuning process. As shown in Figure 1, we demonstrate that our method significantly improves the pose variation and identity restoration of the baseline while also maintaining its performance in pose-irrelevant scenarios. We also found that our method can be extended to the *optimization-based* methods with a single train image.

Our contributions are summarized as:

- For the first time, we unveil the pose-identity entanglement and shed light on a conflict among the visual and textual embeddings.
- Our proposed method effectively resolves the pose bias and identity loss, offering highly diverse and pose-variant subject generation.
- Our method not only is readily applicable to any single-image-based zero-shot customization method, but also further improves the optimization-based methods with a single training image.

Related Works

Text-to-Image Generation. Amidst a proliferation of image synthesis models, diffusion models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020) have demonstrated their strength in producing images with remarkable fidelity and comprehensive mode coverage. Their capacity to generate diverse images has facilitated the integration with large pre-trained language models (Radford et al. 2021). This synergy has given rise to diffusion-based T2I models (Rombach et al. 2022; Balaji et al. 2023; Saharia et al. 2022; Ramesh et al. 2022; Nichol et al. 2022), which can generate high-quality images with strong controllability by the guidance of natural language instructions. Recently, to increase the flexibility using this strong prior, many have combined conditioning embeddings from different modalities, e.g., concatenating text-aligned visual embedding extracted from an image with textual embedding (Sohn et al. 2023; Xiao et al. 2023; Pan et al. 2023; Li, Li, and Hoi 2024; Wei et al. 2023). However, simply combining various embeddings can cause conflict when dealing with different information they contain. Hence, our primary goal is to devise an appropriate methodology for integrating embeddings from different modalities, considering potential conflicts in their inherent information.

Subject-driven Generation. Given a few images of a user-provided subject, subject-driven generation methods (Ruiz

et al. 2023; Gal et al. 2022a; Tewel et al. 2023a; Qiu et al. 2023; Voynov et al. 2023; Kim, Park, and Rhee 2024) aim to generate images containing the subject in various contexts instructed by text guidance. However, per-subject optimization suffers from computation and memory burden, leading to an introduction of zero-shot customization methods (Li, Li, and Hoi 2024; Wei et al. 2023; Chen et al. 2023; Shi et al. 2023; Ma et al. 2023; Jia et al. 2023; Yuan et al. 2023; Zhang et al. 2024; Song et al. 2025; Ye et al. 2023). They involve the mapper that enables the transformation of the input image into text-aligned visual embeddings, bypassing per-subject optimization. We first pin down the pose-identity entanglement in the visual embedding and provide a solution for it, allowing more diverse and more flexible subject-driven generations.

Compositional Generation. Due to the limited size of the textual embeddings, large pre-trained T2I diffusion models suffer from fully compositing complex text descriptions (Liu et al. 2023). Precedent works tackle this problem and offer various solutions e.g., giving an additional segment mask layout for each related prompt as a condition (Kim et al. 2023), composing separate diffusion models where each is encoded with a divided prompt (Liu et al. 2023), and using the perpendicular gradient as a negative prompt guidance (Armandpour et al. 2023). Likewise, we deal with complex prompts consisting of visual and textual embeddings and suggest how to convey the respective information both distinctly and harmoniously.

Preliminaries

In this work, we employ text-to-image latent diffusion model (LDM) (Rombach et al. 2022). The denoising process is implemented in the latent space using the autoencoder structure with the encoder $\mathcal{E}(\cdot)$ and the decoder $\mathcal{D}(\cdot)$. Specifically, an image x is projected to a latent representation $z = \mathcal{E}(x)$, and decoded back to the image space giving $\tilde{x} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(x))$, reconstructing x, i.e., $\tilde{x} \approx x$. Given the pre-trained autoencoder, the latent diffusion model $\epsilon_\theta(z_\tau, \tau, \mathbf{c}); \tau = 1 \cdots T$ is trained with the following objective L_{LDM} where $\mathbf{c}$ represents the contextual embedding of textual/visual condition generally obtained from the pre-trained CLIP encoder (Radford et al. 2021):

$$\mathbb{E}_{x\sim p(x),\epsilon\sim\mathcal{N}(0,I),\mathbf{c},\tau\sim\text{uniform}(1,T)}[\|\epsilon - \epsilon_\theta(z_\tau(x), \tau, \mathbf{c})\|_2^2]. \quad (1)$$

During inference, $z_T \sim \mathcal{N}(0, I)$ is iteratively denoised to the initial representation $z_0(x) = \mathcal{E}(x)$.

Text prompts act as a condition on LDM through the cross-attention mechanism. The latent spatial feature $f \in \mathbb{R}^{l\times h}$ is projected to produce the *query* $Q = f \cdot W_Q \in \mathbb{R}^{l\times d}$ while the text prompts are first encoded into the text embedding $\mathbf{c} \in \mathbb{R}^{l_c\times h_c}$ and projected to yield the *key*, $K = \mathbf{c} \cdot W_K \in \mathbb{R}^{l_c\times d}$, and *value*, $V = \mathbf{c} \cdot W_V \in \mathbb{R}^{l_c\times d}$ where l, l_c, d and h are the spatial sequence length, context sequence length, dimension of key/query/value, and dimension of spatial feature/context respectively. Then the cross-attention is calculated as follows:

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

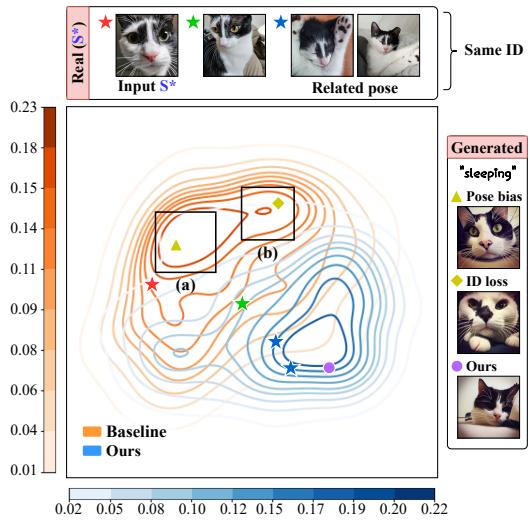

Figure 2: **Dimension reduction (PCA) on the data distribution of generated images.** We visualize real images and generated images with a prompt "A S* sleeping". The baseline tends to generate images with (a) pose bias or (b) identity loss, while our method results in images both identity-conserving and faithful to the text prompt.

Self-attetntion mechanism uses the same Eq. (2) where the key and value are attained from latent spatial features f instead of the contextual embedding $\mathbf{c}$.

Problems: Pose Bias & Identity Loss

Single-image-based zero-shot customization methods (Li, Li, and Hoi 2024; Wei et al. 2023; Chen et al. 2023; Ma et al. 2023; Jia et al. 2023; Yuan et al. 2023; Xiao et al. 2023; Zhang et al. 2024; Song et al. 2025) compose the contextual embedding by concatenating two heterogeneous embeddings; visual $\mathbf{v} \in \mathbb{R}^{M\times h_c}$ and textual $\mathbf{t} \in \mathbb{R}^{N\times h_c}$ ($\mathbf{c} = [\mathbf{v}; \mathbf{t}] = [v_1 \cdots v_M; t_1 \cdots t_N]$) embeddings, where M and N are the number of tokens for the visual and textual embeddings, respectively, and h_c is the dimension of the embedding. Image features of a given subject's image are transformed into the visual embedding using the pre-trained mapper and served as the embedding for a pseudo-word (S*). The visual embedding is then combined with the textual embedding, engaging in the spatial features via cross-attention layers or additional adapters. While users generally desire to give life to their own subject and generate an image of the subject in various poses and actions, a conflict arises within the contextual embedding because the visual embedding already includes the pose information of the subject from the input image. This conflict can lead to two potential issues: either 1) being confined to the specific pose (the visual embedding weakens the textual embedding) or 2) partially losing the subject's identity (the textual embedding impairs the visual embedding).

To better elucidate these two problems, we generated 300 images by BLIP-Diffusion (Li, Li, and Hoi 2024) using a text prompt "A S* sleeping" and extracted their represen-

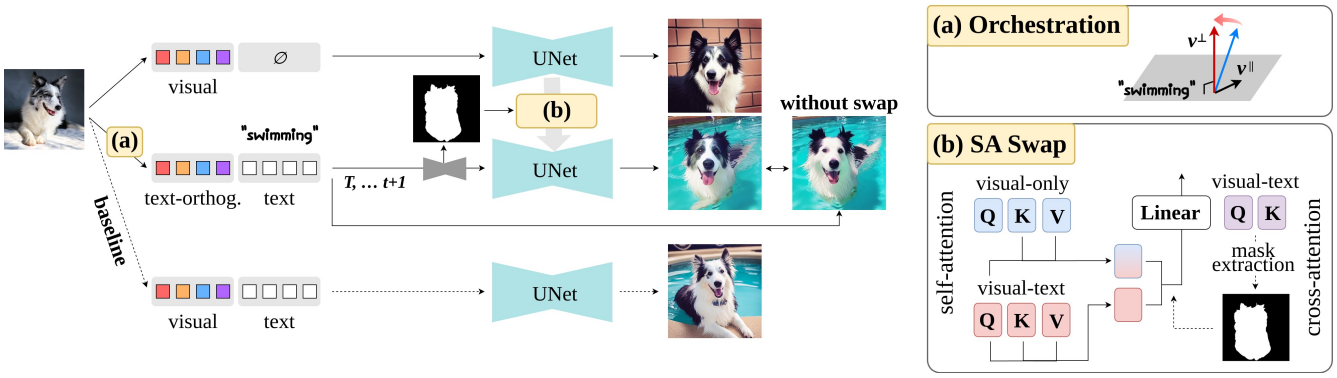

Figure 3: **Overview of our proposed method**. (a) To alleviate the pose bias due to the pose-identity entanglement in the visual embedding, we conduct *Orchestration*, adjusting the visual embedding to be orthogonal to the textual embedding. (b) *Self-attention Swap* obtains self-attention key and value from another denoising process guided by visual-only embedding, which offers the subject's clean identity.

tations from the highest layer of VGG-16 (Simonyan and Zisserman 2014). In Figure 2, the orange contour illustrates images generated by BLIP-D, applying dimension reduction (PCA, $n = 2$) to these representations. Due to the limited number of reference images with the same subject, the actual data distribution is not fully captured in the contour, though individual reference images are marked with stars. Among these, an image directly used for generation is marked with a red star. Images relevant to the desired pose, "sleeping" are marked with blue stars. We found that there are two prominent peaks of BLIP-D generated images. In area (a), generated images are distributed densely around a red star but deviate from blue stars, indicating that they firmly adhere to the input image and its specific pose. On the other hand, area (b) shows a different pattern, where the generated images are shifted away from all the stars as they cannot restore the clean identity of the subject.

Meanwhile, we observe that the images generated by our method, illustrated with blue contours, cover the overall distribution of stars. At the same time, their peak aligns more closely with blue stars, the images with the desired sleeping pose. This result shows that our method can successfully generate the desired pose while effectively preserving the subject's identity.

Methods

Contextual Embedding Orchestration

To resolve the interference between the visual embedding $\mathbf{v}$ and the textual embedding $\mathbf{t}$, we first break down the visual embedding vector v into two components, $v^{||}$ and $v^{\perp}$, where $v^{||}$ resides in the same subspace with the textual embedding $\mathbf{t}$, and $v^{\perp}$ is perpendicular to this subspace. We argue that $v^{||}$ causes the interference with $\mathbf{t}$ when presented concurrently. Meanwhile, $v^{\perp}$ could avoid this interference, and it embodies the essential information about the subject's identity that is orthogonal to the textual indications. Therefore, using $v^{\perp}$ instead of v, we are able to establish the new axes in the visual embedding that interplays more effectively with $\mathbf{t}$. We propose **text-orthogonal visual embedding** $\mathbf{v}^{\perp}$ as follows:

$$v^{\perp} = v - v^{||} = v - \sum_{j=1}^{N} \langle \bar{t}_j, v \rangle \bar{t}_j, \tag{3}$$

$$\bar{t}_j = \text{normal}\left(t_j - \sum_{i=1}^{j-1} \langle \bar{t}_i, t_j \rangle \bar{t}_i \right)$$

where normal$(\cdot)$ means l_2 normalization. $\{\bar{t}_j\}$ are the basis vectors of the textual subspace, obtained by the Gram-Schmidt orthogonalization process.

In detail, we incorporate all text tokens except for articles and the subject's class name in Eq. (3) since they are closely related to pose. For example, when we generate an image of "A S^* playing guitar", the token "guitar" also affects the pose and should be considered, too. The new contextual embedding $\mathbf{c}^{\perp} = [v_1^{\perp} \cdots v_M^{\perp}; t_1 \cdots t_N]$ is then incorporated with latent spatial features via cross-attention. The textual embedding $\mathbf{t}$ effectively guides the diffusion process with alleviated interference from $v^{\perp}$, reducing the effect of pose bias. We illustrate our orchestration in Figure 3(a).

Self-attention Swap

Our orchestration can adeptly resolve conflicts within the contextual embeddings by adjusting the visual embedding. However, the strong entanglement between pose and identity information in the visual embedding can alter the subject's identity during the process of removing the pose information of the subject. To attain the subject's identity, we base our approach on the observation that using the visual-only embedding as the contextual embedding ($\mathbf{c}_{\mathbf{v},\varnothing} = [v_1 \cdots v_M; \varnothing]$) effectively preserves the subject's identity, as it is free from the interference of the textual embedding.

Our objective is to inject the desired features of the visual-only embedding while at the same time maintaining a novel pose of the subject obtained from the text-orthogonal embedding. To this end, we adopt the second denoising process $\{z'_\tau\}_{\tau=1\cdots T}$ where the visual-only embedding is provided as contextual embedding ($\mathbf{c}_{\mathbf{v},\varnothing} = [v_1 \cdots v_M; \varnothing]$). It is distinct

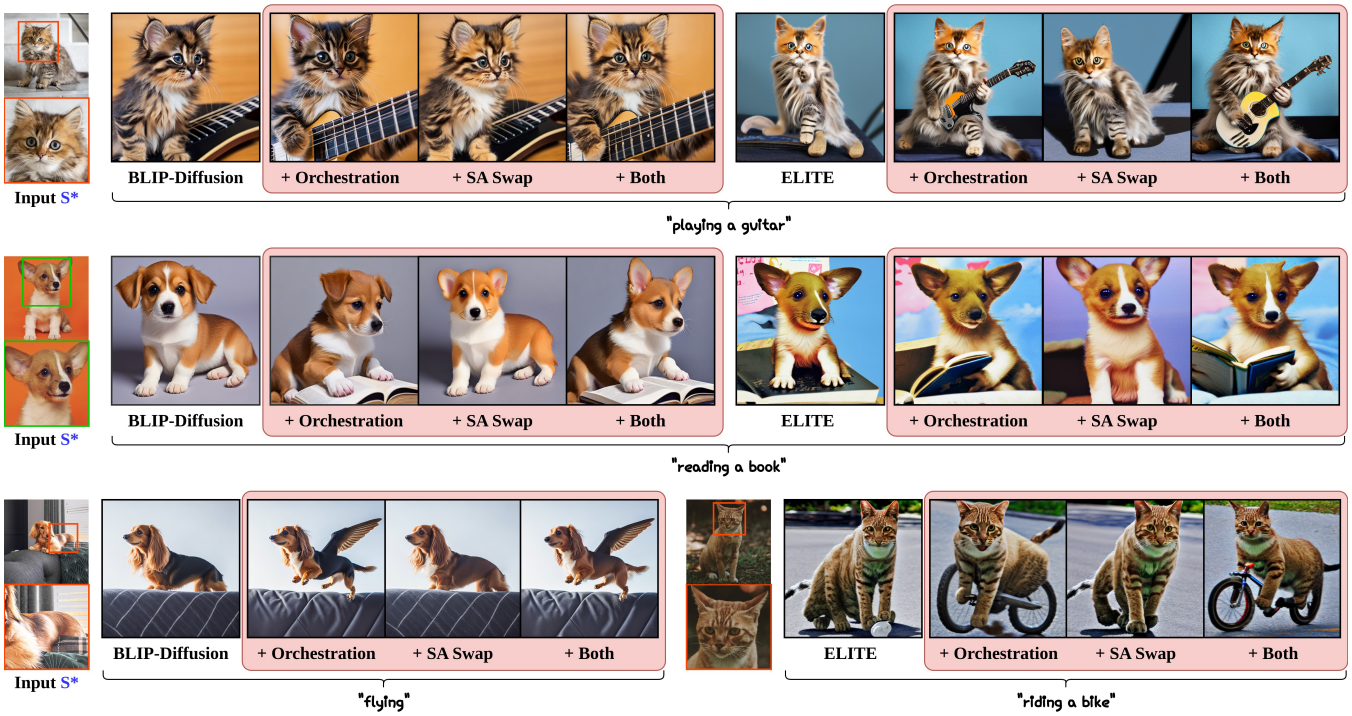

Figure 4: **Comparisons with baseline.** We use the same seed for each pair of images to effectively and progressively demonstrate how our orchestration and self-attention swap improve the baseline.

from the original denoising process $\{z_\tau\}_{\tau=1\cdots T}$ that utilizes the contextual embedding $\mathbf{c}^\perp = [v_1^\perp \cdots v_M^\perp; t_1 \cdots t_N]$. Then, we modify the self-attention layers of $\{z_\tau\}_{\tau=\tau_1\cdots\tau_2}$, to swap the key and value with those from $\{z'_\tau\}_{\tau=\tau_1\cdots\tau_2}$ with timestep hyperparamters τ_1 and τ_2. Our proposed self-attention swap can be formulated as follows:

$$\text{AttnSwap}(z_\tau, z'_\tau) := \text{Attention}(Q_\tau, K'_\tau, V'_\tau). \quad (4)$$

Here, $Q_\tau \in \mathbb{R}^{l\times d}$ is the query from z_τ and $K'_\tau, V'_\tau \in \mathbb{R}^{l\times d}$ are the key and the value from z'_τ where l and d are the spatial sequence length and features dimension, respectively. We generate a novel pose of the subject with the original denoising process while simultaneously incorporating *values* of the clear identity from V' into the location based on the attention map obtained with Q and K', which indicates where the corresponding latent pixels are likely to exist.

Aiming to allow flexibility in the remaining aspects while preserving the subject's identity, we restrict the swaps to latent pixels assigned to the subject. Inspired by (Hertz et al. 2022; Cao et al. 2023), we obtain the binary subject mask, $m \in \{0,1\}^l$, from cross-attention map of token S^* in the original process using a fixed threshold and restore the outputs in the background of the original process as:

$$\text{AttnSwap}(z_\tau, z'_\tau) \odot m + \text{AttnSwap}(z_\tau, z_\tau) \odot (1-m). \quad (5)$$

Here, $\odot$ represents Hadamard product. Note that the second term is the self-attention of the original process z_τ. Figure 3(b) illustrates our self-attention swap process.

Self-attention swap is closely related to the image editing methods that transfer the properties of the source image to the edited image by swapping self-attention (Cao et al. 2023; Tumanyan et al. 2022) or cross-attention (Hertz et al. 2022; Parmar et al. 2023; Couairon et al. 2022). While they edit the source image without changing the other context, our method generates the subject across a range of contexts, accurately incorporating its identity to the proper regions.

Experiments

Datasets. Since prevailing benchmark datasets (Ruiz et al. 2023; Kumari et al. 2023) primarily utilize the generating prompts related to changing the texture or introducing new objects, they often fall short in effectively evaluating the crucial aspect of modifying subject poses. To address this limitation, we have constructed a new dataset, *Deformable Subject Set* (DS set), to effectively assess the model's capability to modify a subject's pose. The DS set comprises 38 live animals from the DreamBooth (Ruiz et al. 2023) and CustomDiffusion (Kumari et al. 2023), along with 11 prompts specifically designed to focus on the deformation of the subjects' poses. Furthermore, we also utilized the *DreamBooth dataset* (DB set) (Ruiz et al. 2023) to evaluate the model's capacity in typical scenarios.

Metrics. Following DreamBooth (Ruiz et al. 2023), we measured the subject fidelity using CLIP-I and DINO-I, and measured text alignment using CLIP-T. For the DS set, which includes object-related action prompts, we found that the newly generated object affects the image-alignment score, for example, by occluding the subject. Therefore, we

Method	Deformable Subject Set					DreamBooth Set		
	CLIP-T(↑)	M-CLIP-I(↑)	M-DINO-I(↑)	CLIP-I(↑)	DINO-I(↑)	CLIP-T(↑)	CLIP-I(↑)	DINO-I(↑)
BLIP-D (Li, Li, and Hoi 2024)	0.262	0.885	0.680	0.835	0.684	0.295	0.812	0.660
*w/*Orchestration	**0.276**	0.883	0.668	0.819	0.663	**0.298**	0.805	0.647
*w/*SA Swap	0.262	**0.886**	**0.689**	**0.837**	**0.694**	0.293	**0.815**	**0.675**
Ours	0.275	**0.886**	0.681	0.821	0.676	0.296	0.809	0.665
ELITE (Wei et al. 2023)	0.285	0.835	0.574	0.751	0.486	0.294	0.792	0.664
*w/*Orchestration	0.292	0.834	0.570	0.754	0.493	**0.295**	0.789	0.658
*w/*SA Swap	0.288	**0.837**	**0.581**	0.754	0.489	0.292	**0.796**	**0.673**
Ours	**0.300**	**0.837**	**0.581**	**0.755**	**0.502**	0.294	0.792	0.670

Table 1: **Quantitative Comparison on Deformable Subject Set and DreamBooth dataset (Ruiz et al. 2023).** 'M-' indicates the metrics using segmentation masks.

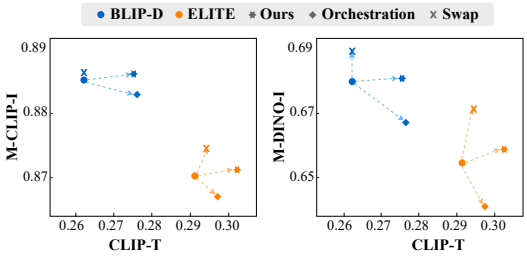

Figure 5: **A Scatter Plot of results from the DS set.** Each component of our method improves its respective axis upon the baselines, with our final method lying on the Pareto front.

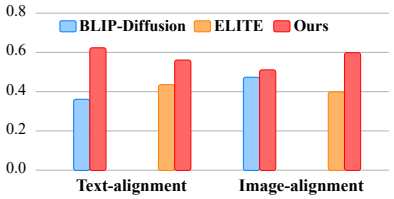

Figure 6: **User Study results**. Our method is steadily preferred over the baselines in both alignments.

additionally measured the masked scores (Avrahami et al. 2023) for CLIP-I and DINO-I, incorporating the subject's segmentation mask for both the input images and generated images. This approach mitigates the impact of new objects added alongside the subject, allowing for a more focused comparison of the subject's identity.

Qualitative Results

We present a comparative analysis of our method with the baselines in Figure 4. While the baselines tend to replicate the subject's pose from the input image, our method effectively modifies the pose aligning with the provided prompt. It is noteworthy that when using object-related prompts, the baselines tend to generate the object without the specified actions, while our method accurately depicts the subject interacting with the given object. Retaining accurate information about the subject from the visual-only embedding, ours also preserves the identity better than the baseline. We use the same seed for images in paired comparisons to validate that each component of our method progressively improves upon the baseline. We provide more various qualitative results, and results from different random seeds, in Appendix.

Quantitative Results

Table 1 and Figure 5 show quantitative analyses. For the DS set, which consists of the prompts necessitating the deformation of the subject's pose, our method significantly improves the text alignment without compromising the image alignment. Each component of our method, orchestration and self-swap, improves text and image alignment, respectively, in line with their objective of pose bias mitigation and identity preservation. We also include additional comparisons with other baselines (Zhang et al. 2024; Song et al. 2025; Ye et al. 2023) in Appendix. Figure 5 illustrates this tendency better with the consistent improvement over the baselines along each axis. Considering both axes involve a trade-off, our combined method becomes closer to the Pareto front, indicating overall improvement. The DB set (Ruiz et al. 2023) consists of prompts unrelated to pose and includes subjects, half of which are non-deformable, making it far from our target. However, quantitative analysis on the DB set in Table 1 confirms that our methodology remains comparable to the existing baseline even in typical scenarios.

User Study

We further evaluate our method through the user study conducted with Amazon Mechanical Turk. Human raters were given a subject's input image, a prompt, and two synthesized images (ours and the baseline) of the subject. They were then asked to select the preferable one for each of the following questions: *(1) Image Alignment*: "Which of the images best reproduces the identity (e.g., item type and details) of the reference item?" *(2) Text Alignment*: "Which of the images is best described by the reference text?". We used 38 animal subjects and 5 to 11 pose-related prompts per subject,

assigning 5 raters for each example. As shown in Figure 6, our method is preferred over the baseline in both text alignment (Ours 62.0% vs. BLIP-Diffusion 38.0%, Ours 56.3% vs. ELITE 43.7%) and image alignment (Ours 51.5% vs. BLIP-Diffusion 48.5%, Ours 60.0% vs. ELITE 40.0%), underscoring our method's ability from a human perspective.

Ablations

Our orchestration eliminates the conflicting elements in the visual embedding *that interfere with the textual embedding.* Then, a subsequent question arises: why don't we directly eliminate the pose information within the visual embedding utilizing the text description of the input image itself? In other words, why not orthogonalize the visual embedding with respect to the text description of the pose found within the input image itself instead of considering an interference with the textual embedding? To investigate the effect of orthogonalization with respect to the embedding vectors, we conduct the following experiment: we adopt the existing image captioning model (Li et al. 2022) and human annotations that describe the pose of the subject in the input image, and transform the visual embedding to be orthogonal to these textual embeddings using Eq. (3). As a result, unalleviated conflict among the contextual embeddings still generates images that are strongly biased toward the input image, leading to lower CLIP-T scores 0.265 and 0.263, respectively than our orchestration (0.276). Resolving this conflict is effective for text alignment and also efficient as it does not require any additional language model or human endeavor.

We also conducted an ablation study on the roles of each component, orchestration and self-attention swap, by progressively applying them. As shown in Table 1 and Figure 4, orchestration makes huge progress on faithfully following the text prompt, while self-attention swap effectively restores the subject's identity. When both are applied, our proposed method successfully generates images that improves both text and image alignments.

Analysis

To verify that the visual embedding is properly adjusted after orchestration, we analyze cross-attention maps for a token corresponding to the visual embedding i.e., a token of the pseudo-word (S*). Cross-attention maps indicate the amount of information conveyed from the tokens to the latent spatial features, therefore, we can obtain insight about the information flows of each token. Figure 7 illustrates that in the baseline (Li, Li, and Hoi 2024), a conflict among the contextual embedding leads the visual embedding to highlight the irrelevant areas. The textual embedding of a pose-related token (<standing>) also shows a similar tendency, resulting in the image noncompliant with a text prompt. Meanwhile, our orchestration effectively resolves the conflict, integrating the embeddings in the proper areas. Quantitatively, we calculate the total sum of the normalized attention score within the subject's segmentation mask (Ren et al. 2024) using the DS set. With our method, attention scores for the visual and textual embedding of the pose-related token are $\times$6.2 and $\times$3.95 more concentrated on the subject compared to the baseline, respectively.

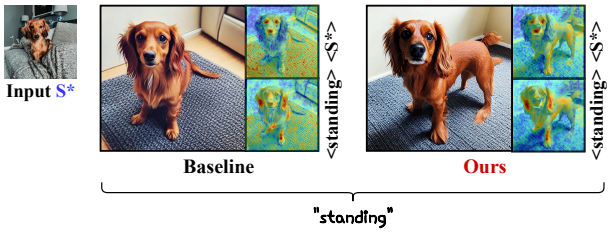

Figure 7: **Visualization of cross-attention map.** (red: high value, blue: low value) Ours injects visual and text information in the proper areas compared to the baseline.

Method	CLIP-T(↑)	M-CLIP-I(↑)	M-DINO-I(↑)
DB	0.301	**0.862**	0.612
Ours	**0.306**	**0.862**	**0.618**
TI	**0.271**	0.845	0.586
Ours	0.264	**0.848**	**0.593**

Table 2: Quantitative results on DS set with optimization-based methods using a single image.

Single-image-based Optimization

Our method addresses a general problem in zero-shot customization. Optimization-based methods are relatively free from these issues, as they update trainable parameters instead of directly using the visual embedding, but they require a lengthy optimization process per subject. However, we found that they also face a similar problem when only a single training image is available, and by applying our method, we can enhance text and image alignment in these methods. Using the DS set, we report the results for Dreambooth (Ruiz et al. 2023) and Textual Inversion (Gal et al. 2022a) in Table 2. For TI, using a single image severely impairs its performance, generating only motion-related objects in a text prompt while omitting the subject. Our method pushes the generated images to restore the subject's identity, leading to the better trade-off between image and text alignment. We provide qualitative results in Appendix.

Conclusion

Limitations. When a set of numerous prompts in different instructions is given at once, our method cannot accurately orthogonalize the visual embedding to each textual embedding, resulting in the images omitting some instructions.

In this paper, we explore notable problems in pose variation tasks when using the single-image-based zero-shot customization methods. We focus on a conflict between the visual and textual embeddings, which affects both embeddings and leads to the *pose bias* and *identity loss*, respectively. We alleviate the conflict by eliminating the interfering elements in the visual embedding while adopting the subject's clean identity from the visual-only embedding. Amidst the proliferation of zero-shot customization methods and the utilization of the visual embedding, our key insight unveils a crucial aspect for achieving a more diverse generation.

Acknowledgments

This work was supported by NRF (2021R1A2C3006659), IITP (RS-2022-II220320, RS-2021-II211343), and KOCCA (RS-2024-00398320), all funded by the Korean Government.

References

Armandpour, M.; Sadeghian, A.; Zheng, H.; Sadeghian, A.; and Zhou, M. 2023. Re-imagine the Negative Prompt Algorithm: Transform 2D Diffusion into 3D, alleviate Janus problem and Beyond. arXiv:2304.04968.

Avrahami, O.; Aberman, K.; Fried, O.; Cohen-Or, D.; and Lischinski, D. 2023. Break-A-Scene: Extracting Multiple Concepts from a Single Image. *arXiv preprint arXiv:2305.16311*.

Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Zhang, Q.; Kreis, K.; Aittala, M.; Aila, T.; Laine, S.; Catanzaro, B.; Karras, T.; and Liu, M.-Y. 2023. eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. arXiv:2211.01324.

Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; and Zheng, Y. 2023. MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 22560–22570.

Chen, L.; Zhao, M.; Liu, Y.; Ding, M.; Song, Y.; Wang, S.; Wang, X.; Yang, H.; Liu, J.; Du, K.; and Zheng, M. 2023. PhotoVerse: Tuning-Free Image Customization with Text-to-Image Diffusion Models. arXiv:2309.05793.

Couairon, G.; Verbeek, J.; Schwenk, H.; and Cord, M. 2022. DiffEdit: Diffusion-based semantic image editing with mask guidance. arXiv:2210.11427.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022a. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. arXiv:2208.01618.

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

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

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. arXiv:2006.11239.

Jia, X.; Zhao, Y.; Chan, K. C. K.; Li, Y.; Zhang, H.; Gong, B.; Hou, T.; Wang, H.; and Su, Y.-C. 2023. Taming Encoder for Zero Fine-tuning Image Customization with Text-to-Image Diffusion Models. arXiv:2304.02642.

Kim, J.; Park, J.; and Rhee, W. 2024. Selectively Informative Description can Reduce Undesired Embedding Entanglements in Text-to-Image Personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8312–8322.

Kim, Y.; Lee, J.; Kim, J.-H.; Ha, J.-W.; and Zhu, J.-Y. 2023. Dense Text-to-Image Generation with Attention Modulation. arXiv:2308.12964.

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

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

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.

Liu, N.; Li, S.; Du, Y.; Torralba, A.; and Tenenbaum, J. B. 2023. Compositional Visual Generation with Composable Diffusion Models. arXiv:2206.01714.

Ma, J.; Liang, J.; Chen, C.; and Lu, H. 2023. Subject-Diffusion:Open Domain Personalized Text-to-Image Generation without Test-time Fine-tuning. arXiv:2307.11410.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. arXiv:2112.10741.

Pan, X.; Dong, L.; Huang, S.; Peng, Z.; Chen, W.; and Wei, F. 2023. Kosmos-g: Generating images in context with multimodal large language models. *arXiv preprint arXiv:2310.02992*.

Parmar, G.; Singh, K. K.; Zhang, R.; Li, Y.; Lu, J.; and Zhu, J.-Y. 2023. Zero-shot Image-to-Image Translation. arXiv:2302.03027.

Qiu, Z.; Liu, W.; Feng, H.; Xue, Y.; Feng, Y.; Liu, Z.; Zhang, D.; Weller, A.; and Schölkopf, B. 2023. Controlling Text-to-Image Diffusion by Orthogonal Finetuning. arXiv:2306.07280.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125.

Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; Zeng, Z.; Zhang, H.; Li, F.; Yang, J.; Li, H.; Jiang, Q.; and Zhang, L. 2024. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. arXiv:2401.14159.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv:2205.11487.

Shi, J.; Xiong, W.; Lin, Z.; and Jung, H. J. 2023. InstantBooth: Personalized Text-to-Image Generation without Test-Time Finetuning. arXiv:2304.03411.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sohl-Dickstein, J.; Weiss, E. A.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. arXiv:1503.03585.

Sohn, K.; Ruiz, N.; Lee, K.; Chin, D. C.; Blok, I.; Chang, H.; Barber, J.; Jiang, L.; Entis, G.; Li, Y.; et al. 2023. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*.

Song, K.; Zhu, Y.; Liu, B.; Yan, Q.; Elgammal, A.; and Yang, X. 2025. Moma: Multimodal llm adapter for fast personalized image generation. In *European Conference on Computer Vision*, 117–132. Springer.

Tewel, Y.; Gal, R.; Chechik, G.; and Atzmon, Y. 2023a. Key-Locked Rank One Editing for Text-to-Image Personalization. arXiv:2305.01644.

Tewel, Y.; Gal, R.; Chechik, G.; and Atzmon, Y. 2023b. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–11.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2022. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. arXiv:2211.12572.

Voynov, A.; Chu, Q.; Cohen-Or, D.; and Aberman, K. 2023. P+: Extended Textual Conditioning in Text-to-Image Generation. arXiv:2303.09522.

Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*.

Xiao, G.; Yin, T.; Freeman, W. T.; Durand, F.; and Han, S. 2023. FastComposer: Tuning-Free Multi-Subject Image Generation with Localized Attention. arXiv:2305.10431.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Yuan, Z.; Cao, M.; Wang, X.; Qi, Z.; Yuan, C.; and Shan, Y. 2023. CustomNet: Zero-shot Object Customization with Variable-Viewpoints in Text-to-Image Diffusion Models. arXiv:2310.19784.

Zhang, Y.; Song, Y.; Liu, J.; Wang, R.; Yu, J.; Tang, H.; Li, H.; Tang, X.; Hu, Y.; Pan, H.; et al. 2024. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8069–8078.