DreamMatcher: Appearance Matching Self-Attention for Semantically-Consistent Text-to-Image Personalization

Jisu Nam*[1], Heesu Kim[2], DongJae Lee[2], Siyoon Jin[1], Seungryong Kim†[1], Seunggyu Chang†[2]

[1]Korea University [2]NAVER Cloud

https://ku-cvlab.github.io/DreamMatcher

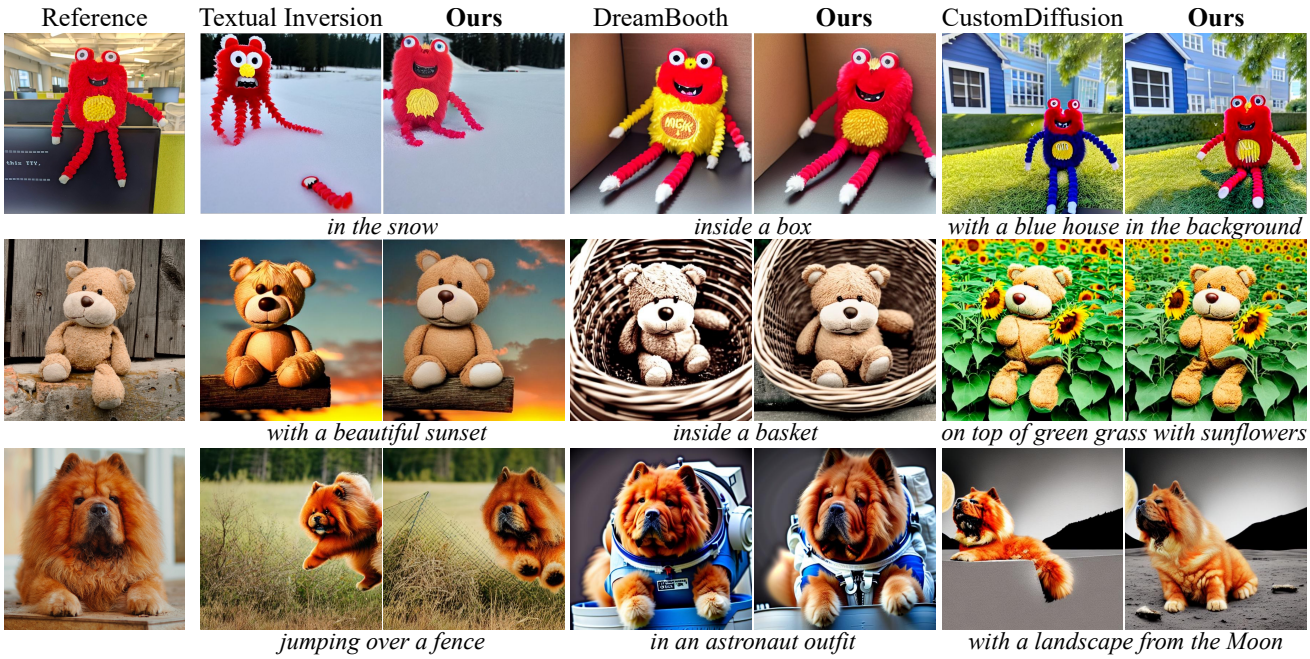


Figure 1. **DreamMatcher enables semantically-consistent text-to-image (T2I) personalization.** Our DreamMatcher is designed to be compatible with any existing T2I personalization models, without requiring additional training or fine-tuning. When integrated with them, DreamMatcher significantly enhances subject appearance, including colors, textures, and shapes, while accurately preserving the target structure as guided by the target prompt.

Abstract

The objective of text-to-image (T2I) personalization is to customize a diffusion model to a user-provided reference concept, generating diverse images of the concept aligned with the target prompts. Conventional methods representing the reference concepts using unique text embeddings often fail to accurately mimic the appearance of the reference. To address this, one solution may be explicitly conditioning the reference images into the target denoising process, known as key-value replacement. However, prior works are constrained to local editing since they disrupt the structure path of the pre-trained T2I model. To overcome this, we propose a novel plug-in method, called DreamMatcher, which reformulates T2I personalization as semantic matching. Specifically, DreamMatcher replaces the target values with reference values aligned by semantic matching, while leaving the structure path unchanged to preserve the versatile capability of pre-trained T2I models for generating diverse structures. We also introduce a semantic-consistent masking strategy to isolate the personalized concept from irrelevant regions introduced by the target prompts. Compatible with existing T2I models, DreamMatcher shows significant improvements in complex scenarios. Intensive analyses demonstrate the effectiveness of our approach.

†Co-corresponding author.
*Work done during an internship at NAVER Cloud.

1. Introduction

The objective of text-to-image (T2I) personalization [17, 32, 44] is to customize T2I diffusion models based on the subject images provided by users. Given a few reference images, they can generate novel renditions of the subject across diverse scenes, poses, and viewpoints, guided by the target prompts.

Conventional approaches [14, 17, 21, 32, 44, 62] for T2I personalization often represent subjects using unique text embeddings [42], by optimizing either the text embedding itself or the parameters of the diffusion model. However, as shown in Figure 1, they often fail to accurately mimic the appearance of subjects, such as colors, textures, and shapes. This is because the text embeddings lack sufficient spatial expressivity to represent the visual appearance of the subject [22, 42]. To overcome this, recent works [8, 10, 18, 29, 34, 48, 53, 63, 65] enhance the expressivity by training T2I models with large-scale datasets, but they require extensive text-image pairs for training.

To address the aforementioned challenges, one solution may be explicitly conditioning the reference images into the target denoising process. Recent subject-driven image editing techniques [4, 9, 11, 28, 31, 37] propose conditioning the reference image through the self-attention module of a denoising U-Net, which is often called key-value replacement. In the self-attention module [25], image features from preceding layers are projected into queries, keys, and values. They are then self-aggregated by an attention operation [61]. Leveraging this mechanism, previous image editing methods [4, 37] replace the keys and values from the target with those from the reference to condition the reference image into the target synthesis process. As noted in [1, 24, 55, 60], we analyze the self-attention module into two distinct paths having different roles for T2I personalization: the query-key similarities form the *structure* path, determining the layout of the generated images, while the values form the *appearance* path, infusing spatial appearance into the image layout.

As demonstrated in Figure 2, our key observation is that the replacement of target keys with reference keys in the self-attention module disrupts the structure path of the pre-trained T2I model. Specifically, an optimal key point for a query point can be unavailable in the replaced reference keys, leading to a sub-optimal matching between target queries and reference keys on the structure path. Consequently, reference appearance is then applied based on this imperfect correspondence. For this reason, prior methods incorporating key and value replacement often fail at generating personalized images with large structural differences, thus being limited to local editing. To resolve this, ViCo [22] incorporates the tuning of a subset of model weights combined with key and value replacement. However, this approach necessitates a distinct tuning process prior to its actual usage.

Figure 2. **Intuition of DreamMatcher:** (a) reference image, (b) disrupted target *structure* path by key-value replacement [4, 9, 11, 28, 31, 37], (c) generated image by (b), (d) target *structure* path in pre-trained T2I model [44], and (e) generated image by DreamMatcher. For visualization, principal component analysis (PCA) [41] is applied to the structure path. Key-value replacement disrupts the target structure, yielding sub-optimal personalized results, whereas DreamMatcher better preserves the target structure, producing high-fidelity subject images aligned with target prompts.

In this paper, we propose a plug-in method dubbed **`DreamMatcher`** that effectively transfers reference appearance while generating diverse structures. DreamMatcher concentrates on the appearance path within the self-attention module for personalization, while leaving the structure path unchanged. However, a simple replacement of values from the target with those from the reference can lead to structure-appearance misalignment. To resolve this, we propose a matching-aware value injection leveraging semantic correspondence to align the reference appearance toward the target structure. Moreover, it is essential to isolate only the matched reference appearance to preserve other structural elements of the target, such as occluding objects or background variations. To this end, we introduce a semantic-consistent masking strategy, ensuring selective incorporation of semantically consistent reference appearances into the target structure. Combined, only the correctly aligned reference appearance is integrated into the target structure through the self-attention module at each time step. However, the estimated reference appearance in early diffusion time steps may lack the fine-grained subject details. To overcome this, we introduce a sampling guidance technique, named semantic matching guidance, to provide rich reference appearance in the middle of the target denoising process.

DreamMatcher is compatible with any existing T2I personalized models without any training or fine-tuning. We show the effectiveness of our method on three different baselines [17, 32, 44]. DreamMatcher achieves state-of-the-art performance compared with existing tuning-free plug-in methods [4, 49, 68] and even a learnable method [22]. As shown in Figure 1, DreamMatcher is effective even in extreme non-rigid personalization scenarios. We further validate the robustness of our method in challenging personalization scenarios. The ablation studies confirm our design choices and emphasize the effectiveness of each component.

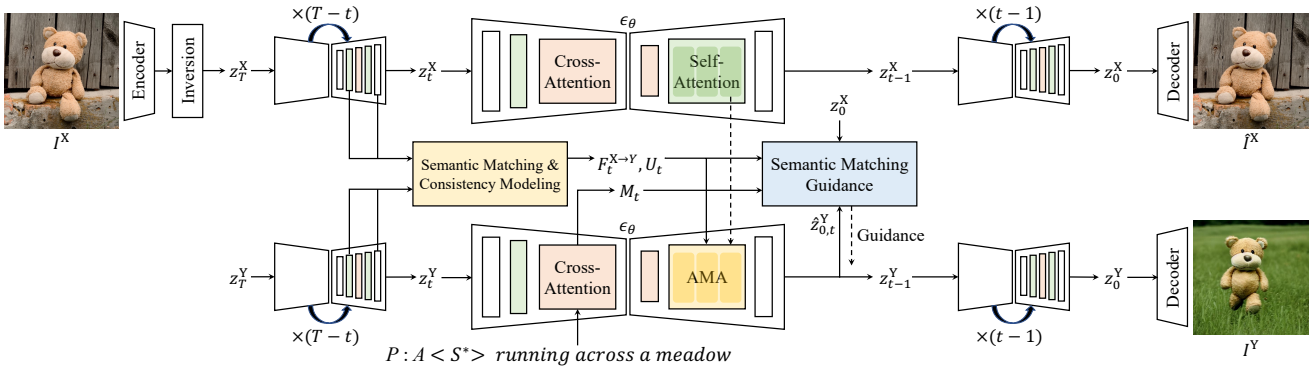


Figure 3. **Overall architecture:** Given a reference image I^X, appearance matching self-attention (AMA) aligns the reference appearance into the fixed target structure in self-attention module of pre-trained personalized model ϵ_θ. This is achieved by explictly leveraging reliable semantic matching from reference to target. Furthermore, semantic matching guidance enhances the fine-grained details of the subject in the generated images.

2. Related Work

Optimization-based T2I Personalization. Given a handful of images, T2I personalization aims to generate new image variations of the given concept that are consistent with the target prompt. Earlier diffusion-based techniques [14, 17, 21, 32, 44, 62] encapsulate the given concept within the textual domain, typically represented by a specific token. Textual Inversion [17] optimizes a textual embedding and synthesizes personalized images by integrating the token with the target prompt. DreamBooth [44] proposes optimizing all parameters of the denoising U-Net based on a specific token and the class category of the subject. Several works [7, 21, 22, 32, 35, 45, 55, 64] focus on optimizing weight subsets or an additional adapter for efficient optimization and better conditioning. For example, CustomDiffusion [32] fine-tunes only the cross-attention layers in the U-Net, while ViCo [22] optimizes an additional image encoder. Despite promising results, the aforementioned approaches often fail to accurately mimic the appearance of the subject.

Training-based T2I Personalization. Several studies [8, 10, 18, 29, 34, 48, 53, 63, 65] have shifted their focus toward training a T2I personalized model with large text-image pairs. For instance, Taming Encoder [29], InstantBooth [48], and FastComposer [65] train an image encoder, while SuTI [10] trains a separate network. While these approaches circumvent fine-tuning issues, they necessitate extensive pre-training with a large-scale dataset.

Plug-in Subject-driven T2I Synthesis. Recent studies [4, 20, 37, 47, 49, 68] aim to achieve subject-driven T2I personalization or non-rigid editing without the need for additional fine-tuning or training. Specifically, MasaCtrl [4] leverages dual-branch pre-trained diffusion models to incorporate image features from the reference branch into the target branch. FreeU [49] proposes reweighting intermediate feature maps from a pre-trained personalized model [44], based on frequency analysis. MagicFusion [68] introduces a noise blending method between a pre-trained diffusion model and a T2I personalized model [44]. DreamMatcher is in alignment with these methods, designed to be compatible with any off-the-shelf T2I personalized models, thereby eliminating additional fine-tuning or training.

3. Preliminary

3.1. Latent Diffusion Models

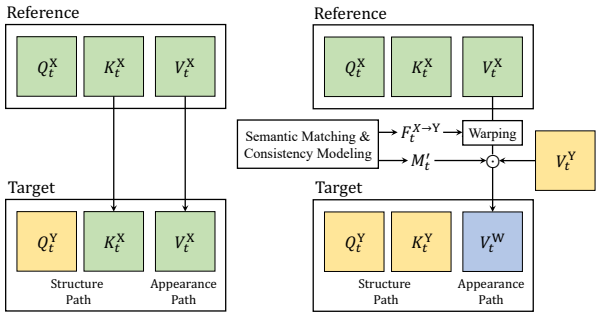
Diffusion models [25, 50] generate desired data samples from Gaussian noise through a gradual denoising process. Latent diffusion models [43] perform this process in the latent space projected by an autoencoder, instead of RGB space. Specifically, an encoder maps an RGB image x_0 into a latent variable z_0 and a decoder then reconstructs it back to x_0. In forward diffusion process, Gaussian noise is gradually added to the latent z_t at each time step t to produce the noisy latent z_{t+1}. In reverse diffusion process, the neural network $\epsilon_\theta(z_t, t)$ denoises z_t to produce z_{t-1} with the time step t. By iteratively sampling z_{t-1}, Gaussian noise z_T is transformed into latent z_0. The denoised z_0 is converted back to x_0 using the decoder. When the condition, e.g., text prompt P, is added, $\epsilon_\theta(z_t, t, P)$ generates latents that are aligned with the text descriptions.

3.2. Self-Attention in Diffusion Models

Diffusion model is often based on a U-Net architecture that includes residual blocks, cross-attention modules, and self-attention modules [25, 43, 50]. The residual block processes the features from preceding layers, the cross-attention module integrates these features with the condition, e.g., text prompt, and the self-attention module aggregates image features themselves through the attention operation.

Specifically, the self-attention module projects the image feature at time step t into queries Q_t, keys K_t, and values V_t. The resulting output from this module is defined by:

$$\mathrm{SA}(Q_t, K_t, V_t) = \mathrm{Softmax}\left(\frac{Q_t K_t^T}{\sqrt{d}}\right) V_t. \quad (1)$$



(a) Key-Value Replacement (b) Appearance Matching Self-Attention

Figure 4. **Comparison between (a) key-value replacement [4, 9, 11, 28, 31, 37] and (b) appearance matching self-attention (AMA):** AMA aligns the reference appearance path toward the fixed target structure path through explicit semantic matching and consistency modeling.

Here, $\mathrm{Softmax}(\cdot)$ is applied over the keys for each query. $Q_t \in \mathbb{R}^{h\times w\times d}$, $K_t \in \mathbb{R}^{h\times w\times d}$, and $V_t \in \mathbb{R}^{h\times w\times d}$ are the projected matrices, where h, w, and d refer to the height, width, and channel dimensions, respectively. As analyzed in [1, 24, 55, 60], we view the self-attention module as two distinct paths: the *structure* and *appearance* paths. More specifically, the structure path is defined by the similarities $\mathrm{Softmax}(Q_t K_t^T/\sqrt{d})$, which controls the spatial arrangement of image elements. The values V_t constitute the appearance path, injecting visual attributes such as colors, textures, and shapes, to each corresponding element within the image.

4. Method

Given a set of n reference images $\mathcal{X} = \{I_n^X\}_{n1}^N$, conventional methods [17, 32, 44] personalize the T2I models $\epsilon_\theta(\cdot)$ with a specific text prompt for the subject (e.g., $\langle S^*\rangle$). In inference, $\epsilon_\theta(\cdot)$ can generate novel scenes from random noises through iterative denoising processes with the subject aligned by the target prompt (e.g., *A* $\langle S^*\rangle$ *in the jungle*). However, they often fail to accurately mimic the subject appearance because text embeddings lack the spatial expressivity to represent the visual attributes of the subject [22, 42]. In this paper, with a set of reference images $\mathcal{X}$ and a target text prompt P, we aim to enhance the subject appearance in the personalized image I^Y, while preserving the detailed target structure directed by the prompt P. DreamMatcher comprises a reference-target dual-branch framework. I^X is inverted to z_T^X via DDIM inversion [50] and then reconstructed to $\hat{I}^X$, while I^Y is generated from a random Gaussian noise z_T^Y guided by P. At each time step, the self-attention module from the reference branch projects image features into queries Q_t^X, K_t^X, and V_t^X, while the target branch produces Q_t^Y, K_t^Y, and V_t^Y. The reference appearance V_t^X is then transferred to the target denoising U-Net through its self-attention module. The overall architecture of DreamMatcher is illustrated in Figure 3.

4.1. Appearance Matching Self-Attention

As illustrated in Figure 4, we propose an appearance matching self-attention (AMA) which manipulates only the appearance path while retaining the pre-trained target structure path, in order to enhance subject expressivity while preserving the target prompt-directed layout.

However, naively swapping the target appearance V_t^Y with that from the reference V_t^X, which reformulates Equation 1, results in structure-appearance misalignment:

$$\mathrm{SA}(Q_t^Y, K_t^Y, V_t^X) = \mathrm{Softmax}\left(\frac{Q_t^Y (K_t^Y)^T}{\sqrt{d}}\right) V_t^X. \quad (2)$$

To solve this, we propose a matching-aware value injection method that leverages semantic matching to accurately align the reference appearance V_t^X with the fixed target structure $\mathrm{Softmax}(Q_t^Y (K_t^Y)^T/\sqrt{d})$. Specifically, AMA warps the reference values V_t^X by the estimated semantic correspondence $F_t^{X\to Y}$ from reference to target, which is a dense displacement field [12, 38, 56, 57, 59] between semantically identical locations in both images. The warped reference values $V_t^{X\to Y}$ are formulated by:

$$V_t^{X\to Y} = \mathcal{W}(V_t^X; F_t^{X\to Y}), \quad (3)$$

where $\mathcal{W}$ represents the warping operation [58].

In addition, it is crucial to isolate only the matched reference appearance and filter out outliers. This is because typical personalization scenarios often involve occlusions, different viewpoints, or background changes that are not present in the reference images, as shown in Figure 1. To achieve this, previous methods [4, 22] use a foreground mask M_t to focus only on the subject foreground and handle background variations. M_t is obtained from the averaged cross-attention map for the subject text prompt (e.g., $\langle S^*\rangle$). With these considerations, Equation 3 can be reformulated as follows:

$$V_t^W = V_t^{X\to Y} \odot M_t + V_t^Y \odot (1 - M_t), \quad (4)$$

where $\odot$ represents Hadamard product [27].

AMA then implants V_t^W into the fixed target structure path through the self-attention module. Equation 2 is reformulated as:

$$\mathrm{AMA}(Q_t^Y, K_t^Y, V_t^W) = \mathrm{Softmax}\left(\frac{Q_t^Y (K_t^Y)^T}{\sqrt{d}}\right) V_t^W. \quad (5)$$

In our framework, we find semantic correspondence between reference and target, aligning with standard semantic matching workflows [12, 13, 26, 38, 56, 57]. Figure 5 provides a detailed schematic of the proposed matching process. In the following, we will explain the process in detail.

Feature Extraction. Classical matching pipelines [12, 13, 26, 38, 56, 57] contain pre-trained feature extractors [6, 23,

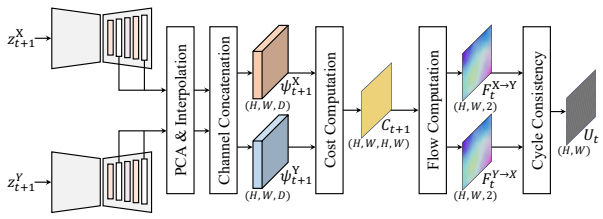


Figure 5. **Semantic matching and consistency modeling:** We leverage internal diffusion features at each time step to find semantic matching $F_t^{X\to Y}$ between reference and target. Additionally, we compute the confidence map of the predicted matches U_t through cycle-consistency.

40] to obtain feature descriptors ψ^X and ψ^Y from image pairs I^X and I^Y. However, finding good features tailored for T2I personalization is not trivial due to the noisy nature of estimated target images in reverse diffusion process, requiring additional fine-tuning of the existing feature extractors. To address this, we focus on the diffusion feature space [54, 67] in the pre-trained T2I model itself to find a semantic matching tailored for T2I personalization.

Let $\epsilon_{\theta,l}(\cdot, t+1)$ denote the output of the l-th decoder layer of the denoising U-Net [25] ϵ_θ at time step $t+1$. Given the latent z_{t+1} with time step $t+1$ and text prompt P as inputs, we extract the feature descriptor $\psi_{t+1,l}$ from the l-th layer of the U-Net decoder. The process is given by:

$$\psi_{t+1,l} = \epsilon_{\theta,l}(z_{t+1},\, t+1,\, P), \tag{6}$$

where we obtain $\psi^X_{t+1,l}$ and $\psi^Y_{t+1,l}$ from z^X_{t+1} and z^Y_{t+1}, respectively. For brevity, we will omit l in the following discussion.

To explore the semantic relationship within the diffusion feature space between reference and target, Figure 6 visualizes the relation between ψ^X_{t+1} and ψ^Y_{t+1} at different time steps using principal component analysis (PCA) [41]. We observe that the foreground subjects share similar semantics, even they have different appearances, as the target image from the pre-trained personalized model often lacks subject expressivity. This observation inspires us to leverage the internal diffusion features to establish semantic matching between estimated reference and target at each time step of sampling phase.

Based on this, we derive $\psi_{t+1} \in \mathbb{R}^{H\times W\times D}$ by combining PCA features from different layers using channel concatenation, where D is the concatenated channel dimension. Detailed analysis and implementation on feature extraction is provided in Appendix E.1.

Flow Computation. Following conventional methods [12, 26, 56, 57, 59], we build the matching cost by calculating the pairwise cosine similarity between feature descriptors for both the reference and target images. For given ψ^X_{t+1} and ψ^Y_{t+1} at time step $t+1$, the matching cost C_{t+1} is computed by taking dot products between all positions in the feature

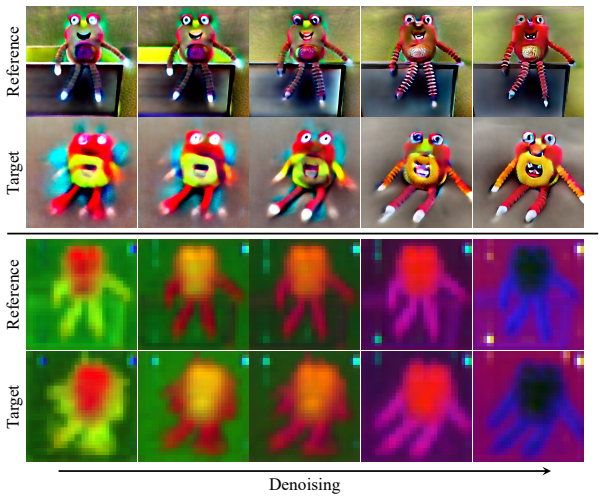


Figure 6. **Diffusion feature visualization:** Upper displays intermediate estimated images of reference and target, with the target generated by DreamBooth [44] using the prompt *A* $\langle S^* \rangle$ *on the beach*. Lower visualizes the three principal components of intermediate diffusion features. The similar semantics share similar colors.

descriptors. This is formulated as:

$$C_{t+1}(i,j) = \frac{\psi^X_{t+1}(i)\cdot\psi^Y_{t+1}(j)}{\|\psi^X_{t+1}(i)\|\|\psi^Y_{t+1}(j)\|}, \tag{7}$$

where $i, j \in [0, H)\times[0, W)$, and $\|\cdot\|$ denotes l-2 normalization.

Subsequently, we derive the dense displacement field from the reference to the target at time step t, denoted as $F_t^{X\to Y} \in \mathbb{R}^{H\times W\times 2}$, using the argmax operation [12] on the matching cost C_{t+1}. Figure 7(c) shows the warped reference image obtained using the predicted correspondence $F_t^{X\to Y}$ between ψ^X_{t+1} and ψ^Y_{t+1} in the middle of the generation process. This demonstrates that the correspondence is established reliably in reverse diffusion process, even in intricate non-rigid target contexts that include large displacements, occlusions, and novel-view synthesis.

4.2. Consistency Modeling

As depicted in Figure 7(d), the forground mask M_t is insufficient to address occlusions and background clutters, (e.g., *a chef outfit* or *a bouquet of flowers*), as these are challenging to distinguish within the cross-attention module.

To compensate for this, we introduce a confidence mask U_t to discard erroneous correspondences, thus preserving detailed target structure. Specifically, we enforce a cycle consistency constraint [30], simply rejecting any correspondence greater than the threshold we set. In other words, we only accept correspondences where a target location x remains consistent when a matched reference location, obtained by $F_t^{Y\to X}$, is re-warped using $F_t^{X\to Y}$. We empirically set the threshold proportional to the target foreground

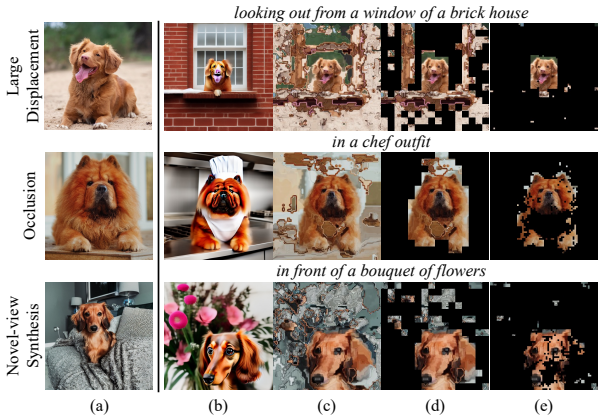


Figure 7. **Correspondence visualization:** (a) Reference image. (b) Estimated target image from DreamBooth [44] at 50% of the reverse diffusion process. (c) Warped reference image based on predicted correspondence $F_t^{X\to Y}$. (d) Warped reference image combined with foreground mask M_t. (e) Warped reference image combined with both M_t and confidence mask U_t.

area. This is formulated by:

$$U_t(x) = \begin{cases} 1, & \text{if } \left\| \mathcal{W}\left(F_t^{Y\to X}; F_t^{X\to Y}\right)(x) \right\| < \gamma\lambda_c, \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where $\|\cdot\|$ denotes a l-2 norm, and $\mathcal{W}$ represents the warping operation [58]. $F_t^{Y\to X}$ indicates the reverse flow field of its forward counterpart, $F_t^{X\to Y}$. γ is a scaling factor designed to be proportional to foreground area, and λ_c is a hyperparameter. More details are available in Appendix E.2.

Finally, we define a semantic-consistent mask M_t' by combining M_t and U_t using the Hadamard product [27], so that M_t coarsely captures the foreground subject, while U_t finely filters out unreliable matches and preserves the fine-grained target context. As shown in Figure 7(e), our network selectively incorporates only the confident matches, effectively addressing intricate non-rigid scenarios.

We now apply a confidence-aware modification to appearance matching self-attention in Equation 4, by replacing M_t with M_t'.

4.3. Semantic Matching Guidance

Our method uses intermediate reference values V_t^X at each time step. However, we observe that in early time steps, these noisy values may lack fine-grained subject details, resulting in suboptimal results. To overcome this, we further introduce a sampling guidance technique, named semantic matching guidance, to provide rich reference semantics in the middle of the target denoising process.

In terms of the score-based generative models [51, 52], the guidance function g steers the target images towards higher likelihoods. The updated direction $\hat{\epsilon}_t$ at time step t is defined as follows [16]:

$$\hat{\epsilon}_t = \epsilon_\theta(z_t, t, P) - \lambda_g \sigma_t \nabla_{z_t} g(z_t, t, P), \tag{9}$$

where λ_g is a hyperparameter that modulates the guidance strength, and σ_t represents the noise schedule parameter at time step t.

We design the guidance function g using z_0^X from DDIM inversion [50], which encapsulates detailed subject representation at the final reverse step $t = 0$. At each time step t, z_0^X is transformed to align with the target structure through $F_t^{X\to Y}$, as follows:

$$z_{0,t}^{X\to Y} = \mathcal{W}(z_0^X; F_t^{X\to Y}). \tag{10}$$

The guidance function g_t at time step t is then defined as the pixel-wise difference between the aligned $z_{0,t}^{X\to Y}$ and the target latent $\hat{z}_{0,t}^Y$ which is calculated by reparametrization trick [50], taking into account the semantic-consistent mask M_t':

$$g_t = \frac{1}{|M_t'|} \sum_{i\in M_t'} \left\| z_{0,t}^{X\to Y}(i) - \hat{z}_{0,t}^Y(i) \right\|, \tag{11}$$

where $\|\cdot\|$ denotes a l-2 norm.

Note that our approach differs from existing methods [2, 16, 37] that provide coarse appearance guidance by calculating the average feature difference between foregrounds. Instead, we leverage confidence-aware semantic correspondence to offer more precise and pixel-wise control.

5. Experiments

5.1. Experimental Settings

Dataset. ViCo [22] gathered an image-prompt dataset from previous works [17, 32, 44], comprising 16 concepts and 31 prompts. We adhered to the ViCo dataset and evaluation settings, testing 8 samples per concept and prompt, for a total of 3,969 images. To further evaluate the robustness of our method in complex non-rigid personalization scenarios, we created a prompt dataset divided into three categories: large displacements, occlusions, and novel-view synthesis. This dataset includes 10 prompts for large displacements and occlusions, and 4 for novel-view synthesis, all created using ChatGPT [39]. The detailed procedure and the prompt list are in the Appendix B.

Baseline and Comparison. DreamMatcher is designed to be compatible with any T2I personalized models. We implemented our method using three baselines: Textual Inversion [17], DreamBooth [44], and CustomDiffusion [32]. We benchmarked DreamMatcher against previous tuning-free plug-in models, FreeU [49] and MagicFusion [68], and also against the optimization-based model, ViCo [22]. Note that additional experiments, including DreamMatcher on Stable Diffusion or DreamMatcher for multiple subject personalization, are provided in Appendix E.

Evaluation Metric. Following previous studies [17, 22, 32, 44], we evaluated subject and prompt fidelity. For subject fidelity, we adopted the CLIP [42] and DINO [5] image

Figure 8. **Qualitative comparison with baselines:** We compare DreamMatcher with three different baselines, Textual Inversion [17], DreamBooth [44], and CustomDiffusion [32].

Method	$I_{\text{DINO}} \uparrow$	$I_{\text{CLIP}} \uparrow$	$T_{\text{CLIP}} \uparrow$
Textual Inversion [17]	0.529	0.762	**0.220**
DreamMatcher	**0.588** (+11.2%)	**0.778** (+2.1%)	0.217 (-1.4%)
DreamBooth [44]	0.638	0.808	**0.237**
DreamMatcher	**0.680** (+6.6%)	**0.821** (+1.6%)	0.231 (-2.5%)
CustomDiffusion [32]	0.667	0.810	0.218
DreamMatcher	**0.700** (+4.9%)	**0.821** (+1.4%)	**0.223** (+2.3%)

Table 1. **Quantitative comparison with different baselines.**

similarity, denoted as I_{CLIP} and I_{DINO}, respectively. For prompt fidelity, we adopted the CLIP image-text similarity T_{CLIP}, comparing visual features of generated images to textual features of their prompts, excluding placeholders. Further details on evaluation metrics are in the Appendix D.1.

User Study. We conducted a user study comparing DreamMatcher to previous works [22, 49, 68]. Participants evaluated the generated images from different methods based on subject and prompt fidelity. 45 users responded to 32 comparative questions, totaling 1440 responses. Samples were chosen randomly from a large, unbiased pool. Additional details on the user study are in Appendix D.2.

5.2. Results

Comparison with Baselines. Table 1 and Figure 8 summarize the quantitative and qualitative comparisons with different baselines. The baselines [17, 32, 44] often lose key visual attributes of the subject such as colors, texture, or shape due to the limited expressivity of text embeddings. In contrast, DreamMatcher significantly outperforms these baselines by a large margin in subject fidelity I_{DINO} and I_{CLIP}, while effectively preserving prompt fidelity T_{CLIP}. As noted in [22, 44], we want to highlight that I_{DINO} better reflects subject expressivity, as it is trained in a self-supervised fashion, thus distinguishing the difference among objects in the same category. Additionally, we wish to note that better prompt fidelity does not always reflect in T_{CLIP}. T_{CLIP} is reported to imperfectly capture text-image alignment and has been replaced by the VQA-based evaluation [19, 66], implying its slight performance drop is negligible. More results are provided in Appendix F.1.

Method	$I_{\text{DINO}} \uparrow$	$I_{\text{CLIP}} \uparrow$	$T_{\text{CLIP}} \uparrow$
MagicFusion [68]	0.632	0.811	0.233
FreeU [49]	0.632	0.806	**0.236**
DreamMatcher	**0.680**	**0.821**	0.231

Table 2. **Quantitative comparison with tuning-free methods.** For this comparison, we used DreamBooth [44] as our baseline.

Method	$I_{\text{DINO}} \uparrow$	$I_{\text{CLIP}} \uparrow$	$T_{\text{CLIP}} \uparrow$
MagicFusion [68]	0.622	0.814	0.235
FreeU [49]	0.611	0.803	**0.242**
DreamMatcher	**0.655**	**0.818**	0.239

Table 3. **Quantitative comparison in challenging dataset.** For this comparison, we used DreamBooth [44] as our baseline.

Method	$I_{\text{DINO}} \uparrow$	$I_{\text{CLIP}} \uparrow$	$T_{\text{CLIP}} \uparrow$
ViCo [22]	0.643	0.816	**0.228**
DreamMatcher	**0.700**	**0.821**	0.223

Table 4. **Comparison with optimization-based method.** For this comparison, we used CustomDiffusion [32] as our baseline.

Comparison with Plug-in Models. We compared DreamMatcher against previous tuning-free plug-in methods, FreeU [49] and MagicFusion [68]. Both methods demonstrated their effectiveness when plugged into DreamBooth [44]. For a fair comparison, we evaluated DreamMatcher using DreamBooth as a baseline. As shown in Table 2 and Figure 9, DreamMatcher notably outperforms these methods in subject fidelity, maintaining comparable prompt fidelity. The effectiveness of our method is also evident in Table 3, displaying quantitative results in challenging non-rigid personalization scenarios. This highlights the importance of semantic matching for robust performance in complex real-world personalization applications.

Comparison with Optimization-based Models. We further evaluated DreamMatcher against the optimization-based model, ViCo [22], which fine-tunes an image adapter with 51.3M parameters. For a balanced comparison, we compared ViCo with DreamMatcher combined with CustomDiffusion [32], configured with a similar count of train-

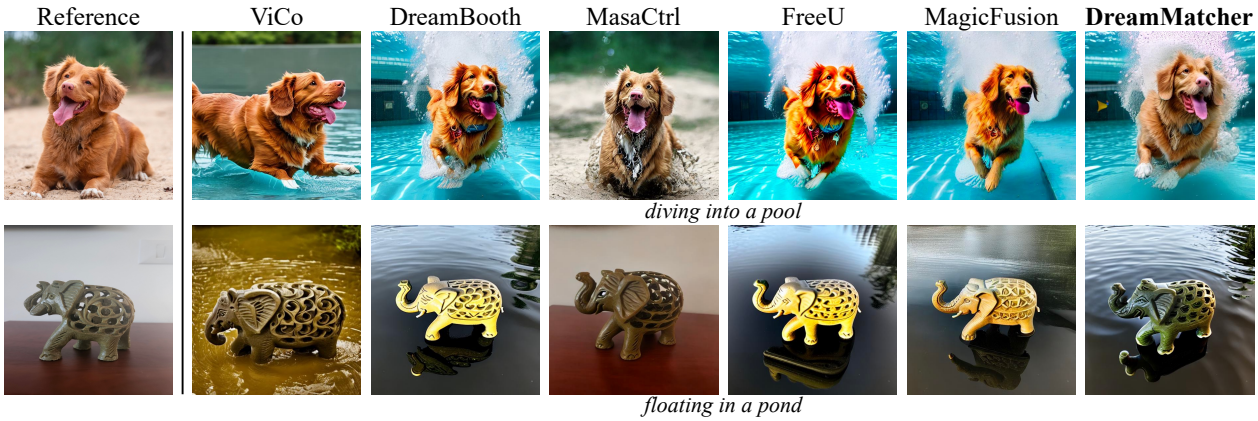


Figure 9. **Qualitative comparison with previous works** [4, 22, 44, 49, 68]**:** For this comparison, DreamBooth [44] was used as the baseline of MasaCtrl, FreeU, MagicFusion, and DreamMatcher.

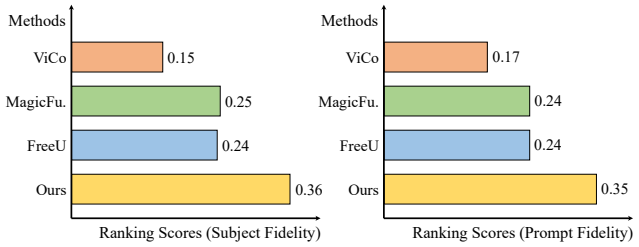


Figure 10. **User study.**

able parameters (57.1M). Table 4 shows DreamMatcher notably surpasses ViCo in all subject fidelity metrics, without requiring extra fine-tuning. Figure 9 provides the qualitative comparison. More results are provided in Appendix F.2.

User Study. We also present the user study results in Figure 10, where DreamMatcher significantly surpasses all other methods in both subject and prompt fidelity. Further details are provided in Appendix D.2.

5.3. Ablation Study

In Figure 11 and Table 5, we demonstrate the effectiveness of each component in our framework. (b) and (I) present the results of the baseline, while (II) shows the results of key-value replacement, which fails to preserve the target structure and generates a static subject image. (c) and (III) display AMA using predicted correspondence, enhancing subject fidelity compared to (b) and (I), but drastically reducing prompt fidelity, as it could not filter out unreliable matches. This is addressed in (d) and (IV), which highlight the effectiveness of the semantic-consistent mask in significantly improving prompt fidelity, up to the baseline (I). Finally, the comparison between (d) and (e) demonstrate that semantic-matching guidance improves subject expressivity with minimal sacrifice in target structure, which is further evidenced by (V). More analyses, including a user study comparing DreamMatcher and MasaCtrl, are in Appendix E.

Figure 11. **Component analysis:** (a) reference image, (b) generated image by DreamBooth [44], (c) with proposed semantic matching, (d) further combined with semantic-consistent mask, and (e) further combined with semantic matching guidance.

	Component	$I_{\text{DINO}} \uparrow$	$I_{\text{CLIP}} \uparrow$	$T_{\text{CLIP}} \uparrow$
(I)	Baseline (DreamBooth [44])	0.638	0.808	0.237
(II)	(I) + Key-Value Replacement (MasaCtrl [4])	0.728	0.854	0.201
(III)	(I) + Semantic Matching	0.683	0.830	0.201
(IV)	(III) + Semantic-Consistent Mask (AMA)	0.676	0.818	0.232
(V)	(IV) + Semantic Matching Guid. (Ours)	0.680	0.821	0.231

Table 5. **Component analysis.** For this analysis, we used DreamBooth [44] for the baseline.

6. Conclusion

We present DreamMatcher, a tuning-free plug-in for text-to-image (T2I) personalization. DreamMatcher enhances appearance resemblance in personalized images by providing semantically aligned visual conditions, leveraging the generative capabilities of the self-attention module within pre-trained T2I personalized models. DreamMatcher pioneers the significance of semantically aligned visual conditioning in personalization, offering an effective solution within the attention framework. Experiments show that DreamMatcher enhances the personalization capabilities of existing T2I models, outperforming previous tuning-free plug-ins, even in complex scenarios.

7. Acknowledgements

This research was supported by the MSIT, Korea (IITP-2024-2020-0-01819, ICT Creative Consilience Program, No.2021-0-02068, Artificial Intelligence Innovation Hub).

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2, 4

[2] 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*, pages 843–852, 2023. 6

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. A.6

[4] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. *arXiv preprint arXiv:2304.08465*, 2023. 2, 3, 4, 8, A.4, A.5, A.6, A.10, A.15, A.16

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 6, A.2

[6] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014. 4

[7] Hong Chen, Yipeng Zhang, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. Disenbooth: Disentangled parameter-efficient tuning for subject-driven text-to-image generation. *arXiv preprint arXiv:2305.03374*, 2023. 3

[8] Li Chen, Mengyi Zhao, Yiheng Liu, Mingxu Ding, Yangyang Song, Shizun Wang, Xu Wang, Hao Yang, Jing Liu, Kang Du, et al. Photoverse: Tuning-free image customization with text-to-image diffusion models. *arXiv preprint arXiv:2309.05793*, 2023. 2, 3, A.5

[9] Songyan Chen and Jiancheng Huang. Fec: Three finetuning-free methods to enhance consistency for real image editing. *arXiv preprint arXiv:2309.14934*, 2023. 2, 4, A.4

[10] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186*, 2023. 2, 3, A.5

[11] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023. 2, 4, A.4

[12] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. Cats: Cost aggregation transformers for visual correspondence. *Advances in Neural Information Processing Systems*, 34:9011–9023, 2021. 4, 5

[13] Seokju Cho, Sunghwan Hong, and Seungryong Kim. Cats++: Boosting cost aggregation with convolutions and transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 4

[14] Ziyi Dong, Pengxu Wei, and Liang Lin. Dreamartist: Towards controllable one-shot text-to-image generation via contrastive prompt-tuning. *arXiv preprint arXiv:2211.11337*, 2022. 2, 3

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. A.3

[16] Dave Epstein, Allan Jabri, Ben Poole, Alexei A Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *arXiv preprint arXiv:2306.00986*, 2023. 6

[17] 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. 2, 3, 4, 6, 7, A.2, A.3, A.5, A.6, A.8, A.13, A.14

[18] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Designing an encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2302.12228*, 2023. 2, 3, A.5

[19] Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *arXiv preprint arXiv:2310.11513*, 2023. 7

[20] Jing Gu, Yilin Wang, Nanxuan Zhao, Tsu-Jui Fu, Wei Xiong, Qing Liu, Zhifei Zhang, He Zhang, Jianming Zhang, HyunJoon Jung, et al. Photoswap: Personalized subject swapping in images. *arXiv preprint arXiv:2305.18286*, 2023. 3

[21] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023. 2, 3

[22] Shaozhe Hao, Kai Han, Shihao Zhao, and Kwan-Yee K Wong. Vico: Detail-preserving visual condition for personalized text-to-image generation. *arXiv preprint arXiv:2306.00971*, 2023. 2, 3, 4, 6, 7, 8, A.2, A.3, A.6, A.9, A.15, A.16

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[24] 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. 2, 4, A.6

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3, 5

[26] Sunghwan Hong, Jisu Nam, Seokju Cho, Susung Hong, Sangryul Jeon, Dongbo Min, and Seungryong Kim. Neural matching fields: Implicit representation of matching fields

for visual correspondence. *Advances in Neural Information Processing Systems*, 35:13512–13526, 2022. 4, 5

[27] Roger A Horn. The hadamard product. In *Proc. Symp. Appl. Math*, pages 87–169, 1990. 4, 6

[28] Jiancheng Huang, Yifan Liu, Jin Qin, and Shifeng Chen. Kv inversion: Kv embeddings learning for text-conditioned real image action editing. *arXiv preprint arXiv:2309.16608*, 2023. 2, 4, A.4

[29] Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*, 2023. 2, 3, A.5

[30] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6207–6217, 2021. 5

[31] Anant Khandelwal. Infusion: Inject and attention fusion for multi concept zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3017–3026, 2023. 2, 4, A.4

[32] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 2, 3, 4, 6, 7, A.2, A.3, A.5, A.6, A.8, A.13, A.14

[33] Jason Lee, Kyunghyun Cho, and Douwe Kiela. Countering language drift via visual grounding. *arXiv preprint arXiv:1909.04499*, 2019. A.2

[34] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023. 2, 3, A.5

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

[36] Yuchen Lu, Soumye Singhal, Florian Strub, Aaron Courville, and Olivier Pietquin. Countering language drift with seeded iterated learning. In *International Conference on Machine Learning*, pages 6437–6447. PMLR, 2020. A.2

[37] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. *arXiv preprint arXiv:2307.02421*, 2023. 2, 3, 4, 6, A.3, A.4

[38] Jisu Nam, Gyuseong Lee, Sunwoo Kim, Hyeonsu Kim, Hyoungwon Cho, Seyeon Kim, and Seungryong Kim. Diffmatch: Diffusion model for dense matching. *arXiv preprint arXiv:2305.19094*, 2023. 4

[39] OpenAI. Gpt-4 technical report, 2023. 6, A.2

[40] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5

[41] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901. 2, 5, A.3, A.5

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 4, 6, A.2

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

[44] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2, 3, 4, 5, 6, 7, 8, A.2, A.3, A.4, A.5, A.6, A.8, A.13, A.14, A.15, A.16

[45] 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. *arXiv preprint arXiv:2307.06949*, 2023. 3

[46] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. A.2

[47] Junyoung Seo, Gyuseong Lee, Seokju Cho, Jiyoung Lee, and Seungryong Kim. Midms: Matching interleaved diffusion models for exemplar-based image translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2191–2199, 2023. 3

[48] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 2, 3, A.5

[49] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. *arXiv preprint arXiv:2309.11497*, 2023. 2, 3, 6, 7, 8, A.2, A.3, A.6, A.9, A.15, A.16

[50] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 4, 6, A.2

[51] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 6

[52] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 6

[53] Yu-Chuan Su, Kelvin CK Chan, Yandong Li, Yang Zhao, Han Zhang, Boqing Gong, Huisheng Wang, and Xuhui Jia.

Identity encoder for personalized diffusion. *arXiv preprint arXiv:2304.07429*, 2023. 2, 3, A.5

[54] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881*, 2023. 5, A.3

[55] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2, 3, 4

[56] Prune Truong, Martin Danelljan, Luc V Gool, and Radu Timofte. Gocor: Bringing globally optimized correspondence volumes into your neural network. *Advances in Neural Information Processing Systems*, 33:14278–14290, 2020. 4, 5

[57] Prune Truong, Martin Danelljan, and Radu Timofte. Glu-net: Global-local universal network for dense flow and correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6258–6268, 2020. 4, 5

[58] Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Warp consistency for unsupervised learning of dense correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10346–10356, 2021. 4, 6

[59] Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. Pdc-net+: Enhanced probabilistic dense correspondence network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 4, 5

[60] 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*, pages 1921–1930, 2023. 2, 4, A.6

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

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

[63] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 2, 3, A.5

[64] Chendong Xiang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. A closer look at parameter-efficient tuning in diffusion models. *arXiv preprint arXiv:2303.18181*, 2023. 3

[65] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 2, 3, A.5

[66] Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What you see is what you read? improving text-image alignment evaluation. *arXiv preprint arXiv:2305.10400*, 2023. 7

[67] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *arXiv preprint arXiv:2305.15347*, 2023. 5, A.3

[68] Jing Zhao, Heliang Zheng, Chaoyue Wang, Long Lan, and Wenjing Yang. Magicfusion: Boosting text-to-image generation performance by fusing diffusion models. *arXiv preprint arXiv:2303.13126*, 2023. 2, 3, 6, 7, 8, A.2, A.3, A.6, A.9, A.15, A.16