DynASyn: Multi-Subject Personalization Enabling Dynamic Action Synthesis

Yongjin Choi, Chanhun Park, Seung Jun Baek[*]

Department of Computer Science, Korea University, Seoul, Korea
{dydwls8445, cksgns2010, sjbaek}@korea.ac.kr

Abstract

Recent advances in text-to-image diffusion models spurred research on personalization, i.e., a customized image synthesis, of subjects within reference images. Although existing personalization methods are able to alter the subjects' positions or to personalize multiple subjects simultaneously, they often struggle to modify the behaviors of subjects or their dynamic interactions. The difficulty is attributable to overfitting to reference images, which worsens if only a single reference image is available. We propose DynASyn, an effective multi-subject personalization from a single reference image addressing these challenges. DynASyn preserves the subject identity in the personalization process by aligning concept-based priors with subject appearances and actions. This is achieved by regularizing the attention maps between the subject token and images through concept-based priors. In addition, we propose concept-based prompt-and-image augmentation for an enhanced trade-off between identity preservation and action diversity. We adopt an SDE-based editing guided by augmented prompts to generate diverse appearances and actions while maintaining identity consistency in the augmented images. Experiments show that DynASyn is capable of synthesizing highly realistic images of subjects with novel contexts and dynamic interactions with the surroundings, and outperforms baseline methods in both quantitative and qualitative aspects.

Introduction

Recent advances in text-to-image (T2I) generative models (Rombach et al. 2022; Ramesh et al. 2022; Saharia et al. 2022) have enabled the rendition of highly realistic and creative images. These models are trained on large datasets of image-text pairs like LAION (Schuhmann et al. 2022), allowing them to generate novel images from text prompts. In particular, there has been growing interest on T2I personalization (Gal et al. 2022; Ruiz et al. 2023a). The T2I personalization is a task of generating variations of user-provided images in novel contexts by modifying aspects such as pose, action, color, and interactions between subjects.

The studies on T2I personalization initially focused on the synthesis regarding a single subject given multiple images containing the subject (Gal et al. 2022; Ruiz et al. 2023a) by finetuning the text embedding representing the subject. Subsequent works (Wu et al. 2023; Wang et al. 2024a; Hua et al. 2023) explored personalization of a single subject based on a single image by using an additional image encoder. However, they had difficulties in varying the actions of personalized objects or required additional constraints like depth maps or skeletons. Recent works (Avrahami et al. 2023; Kim et al. 2024; Matsuda et al. 2024) proposed personalization of multiple subjects from a single image, enabling more convenient and diverse applications. However, they struggled to modulate diverse actions (Avrahami et al. 2023) or needed explicit spatial guidance on subject regions when generating additional images (Kim et al. 2024; Matsuda et al. 2024).

We aim to achieve multi-subject personalization from a single image while avoiding overfitting by introducing regularization through attention maps guided by the subject's concept-based prior. These priors, including class information, capture typical behaviors and enhance generalization. Unlike existing approaches that use MSE loss between segmentation masks and attention maps to isolate identities (Avrahami et al. 2023; Xiao et al. 2023; Wang et al. 2024b), which ignore conceptual priors, our method employs loss functions on attention maps informed by subject concepts.

Furthermore, we implement concept-based prompt and image augmentation to encompass the subject's descriptive attributes and actions. Specifically, our Guided SDE Augmentation (GSA), inspired by SDEdit (Meng et al. 2021), directs a T2I model with augmented prompts to generate images that balance identity preservation and action diversity. These augmentations enable diverse appearances, actions, and dynamic interactions among subjects. Experiments show that DynASyn produces images that are quantitatively and qualitatively superior to prior state-of-the-art methods. Our main contributions are summarized as follows. (1) We propose DynASyn for personalizing multiple subjects from a single image, which aligns the concept-based priors with the subject appearances and actions to mitigate overfitting. (2) We propose concept-based attention regularization and prompt-and-image augmentation for an effective alignment with concept priors. (3) Experiments on diverse datasets and prompts demonstrate that DynASyn achieves state-of-the-art personalization capabilities of synthesis for novel contexts and dynamic actions for the subjects.
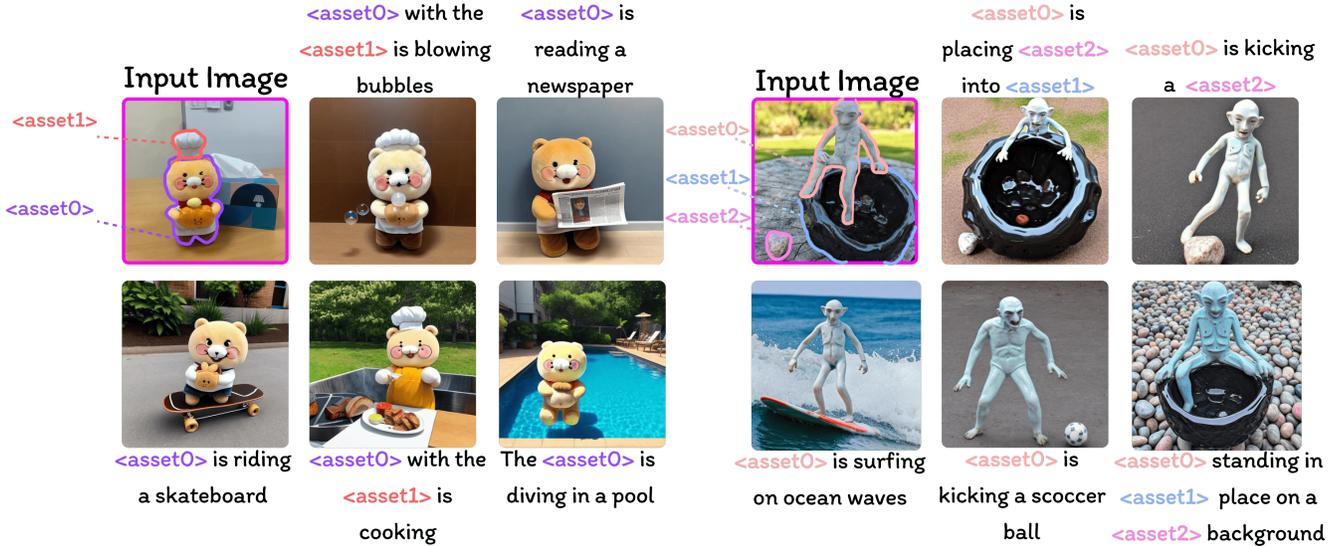
---

[*]Corresponding Author

Figure 1: Personalization outputs from the proposed method, DynASyn. When provided with a single image containing multiple subjects, each subject can be trained into placeholders denoted as <asset>. DynASyn is capable of synthesizing diverse types of novel poses and dynamic actions of the subjects from the text prompts by avoiding overfitting to the reference image.

Related Work

Text-to-image Synthesis and Editing The advent of large-scale image-text datasets (Schuhmann et al. 2022) and computing resources have spurred rapid innovations in text-to-image (T2I) synthesis models. Beginning with GAN based T2I model (Kang et al. 2023; Li et al. 2022), the field transitioned to transformer-based architectures (Ding et al. 2022; Ramesh et al. 2021). More recently, diffusion-based models (Rombach et al. 2022; Ramesh et al. 2022; Yu et al. 2022; Saharia et al. 2022) have demonstrated remarkable progress. Diffusion models for T2I learn to progressively denoise images or latent representations from Gaussian noise, overcoming many limitations of existing generative models based on GANs. When conditioned on text, they can produce strikingly realistic and diverse images. However, a major shortcoming is the lack of control over consistent subject generation or steering outputs towards desired target images.

Seeking more control beyond text-conditional image generation, recent work explores editing diffusion models by removing, inserting, or modifying parts of a given image. Some approaches edit the cross-attention map to alter sections of the image or replace subjects (Chefer et al. 2023; Feng et al. 2022). Others incorporate additional signals like bounding boxes to control spatial regions (Chen, Laina, and Vedaldi 2024; Li et al. 2023). Recently, combining diffusion models with large language models (LLMs) in prompt-to-prompt frameworks (Hertz et al. 2022) has enabled new pipeline designs for controllable editing.

T2I Personalization T2I personalization involves adapting models to generate images of a particular object given a few reference photos, conditioned on novel text prompts. Textual Inversion (TI) (Gal et al. 2022) maps a set of 4-6 images of a subject to a placeholder token, training the text embedding to point to that specific subject. In TI, the small number of parameters updated during training can limit the identity preservation. DreamBooth (DB) (Ruiz et al. 2023a) follows a similar paradigm but also finetunes the parameters of the UNet to strengthen the identity preservation. However, DB can overfit to the reference images, which limits the generalization capabilities. The research on simultaneous personalization of multiple subjects initially focused on human faces. Those works trained additional face encoders (Xiao et al. 2023; Ruiz et al. 2023b; Wei et al. 2024) to disambiguate identities or used loss functions between segmentation masks and attention maps of different subjects (Xiao et al. 2023; Wei et al. 2024). CustomDiffusion (Kumari et al. 2023) and Perfusion (Tewel et al. 2023) enable multi-subject personalization from a few images by solely updating the key and value matrices in the cross-attention layer. Recent methods (Gal et al. 2023; Wei et al. 2023; Jia et al. 2023; Li, Li, and Hoi 2024) enable multi-subject personalization from a single image, but they require training additional encoders similar to the face personalization, which demands substantial computing resources and data. Break-a-Scene (Avrahami et al. 2023) was able to personalize multiple subjects present in a single image. The method distinguished subjects using losses between segmentation masks and attention maps per subject, and trained place-holder tokens via an objective function combining TI and DB. However, the attention loss between subject attention maps and masks in Break-a-Scene may cause overfitting to fine-grained appearance details of the reference image. As a result, generations conditioned on new text tend to similar to the input image or to have difficulties in expressing novel poses or actions of the subject.

Preliminaries

Text-to-Image Diffusion Models

Text-to-image (T2I) models (Rombach et al. 2022; Ramesh et al. 2022; Saharia et al. 2022) trained on large-scale image-text datasets synthesize high-fidelity images which accurately reflect the visual concepts expressed in textual descriptions. During training, the model learns to predict the text-conditional noise residual between the original and noisy images. For sampling, the model progressively denoises random images conditioned on the text embedding to generate the final image. Given a text encoder, γ_θ, the training objective is given by

$$\mathbb{E}_{z,c,\epsilon\sim\mathcal{N}(0,1),t}[||\epsilon - \epsilon_\theta(\gamma_\theta(c), z_t, t)||_2^2] \quad (1)$$

where t is the timestep, z_t is the noisy latent, and ϵ_θ is the denoising model. In T2I personalization, the model is trained on text prompts where a placeholder token denoted by `<asset>` represents the subject image. The model optimizes the embedding of placeholder tokens guided by the reconstruction loss. During training, text embeddings are incorporated as conditions via cross-attention with the spatial features of the UNet architecture. Specifically, the queries Q come from an intermediate spatial latent of the UNet, and the keys K and values V are derived from the output of the text encoder network $\gamma_\theta(c)$. The resulting cross-attention map is $A_t = \text{softmax}(QK^T/\sqrt{d})$ where d represents the dimension of the projected K and Q. In $A_t \in \mathbb{R}^{m \times m \times N}$, m is the spatial dimension of the attention map. A_t plays a pivotal role in our method, and its use will be explained in the sequel. We adopt Stable Diffusion v2.1 (Rombach et al. 2022) as the main T2I model.

Masked Diffusion

Recent approaches to multi-subject personalization utilized the segmentation masks of subjects during training (Avrahami et al. 2023; Xiao et al. 2023; Wei et al. 2024). The reconstruction loss is applied to the region activated by the segmentation mask of each subject. We also take this approach, where the subjects at each training step are randomly selected via *union sampling* (Avrahami et al. 2023) as follows. Suppose there are a total of N subjects, and let $\mathcal{I} = \{1, \ldots, N\}$ denote the set of subject indices. A non-empty random subset $s \subseteq \mathcal{I}$ is sampled at each training step. Let M_i denote the mask of subject i and $M_s = \bigcup_{i \in s} M_i$ denote the union of the masks. The *masked diffusion loss* $\mathcal{L}_{\text{MD}}$ is defined as

$$\mathcal{L}_{\text{MD}} = \mathbb{E}_{z,s,\epsilon\sim\mathcal{N}(0,1),t} \left[||\epsilon \odot M_s - \epsilon_\theta(z_t, t, \gamma_\theta(c)) \odot M_s||_2^2 \right] \quad (1)$$

which is calculated by applying the subjects' masks to the noisy image. In addition, the masked loss can be applied to the cross-attention map between the placeholder token and images obtained through text-conditioning. For subject $i \in s$ and time t, let $A_{t,i}(\gamma, z)$ denote the cross-attention map between the i-th token embedding from γ and the latent z at t. The masked attention loss $\mathcal{L}_{\text{M2A}}$ is given by

$$\mathcal{L}_{\text{M2A}} = \mathbb{E}_{z,s,t} \left[\sum_{i \in s} ||A_{t,i}(\gamma_\theta(c), z_t) - M_i||_2^2 \right]. \quad (2)$$

A benefit of masked diffusion with cross-attention is that, the identity of subject can be well-preserved by guiding the alignment between the subject placeholder and images through attention. However, at the same time, $\mathcal{L}_{\text{M2A}}$ encourages the model to attend uniformly to all areas within the mask. This can cause the model to overfit to mask outlines and geometries of subjects, which limits the generation of novel poses or actions for personalization. Thus, there exists a tension between preserving identity and overfitting to input seed images.

Proposed Method

We aim to personalize *multiple* subjects in a *single* image by learning to disentangle the subjects and to generate novel concepts, actions or interactions involving the subjects. A key to successful personalization is on how to properly integrate the subject identity and class-conditional priors of the subject. A vast knowledge of such priors can be found from large-scale T2I models, e.g., Stable Diffusion. The cross-attention map between the text describing the subject class and the associated images generated by T2I models contains rich information on the class-conditional image priors. We propose to utilize those attention maps to infuse class-specific concepts into subject placeholders so as to facilitate generating novel appearances and actions consistent with the class priors.

Class-Specific Attention Regularization

We first focus on aligning subject *appearances* with class-specific priors to mitigate overfitting issues. For optimizing the placeholder embeddings associated with the input images, a neutral prompt for the subject is used. For example, 'a photo of <asset0> and <asset1>' is input to the text encoder, where <asset0> and <asset1> refer to the panda and bowl in the input image, e.g., see Fig. 2 (a). Next, we create another prompt with placeholders in personalization prompts replaced by class names of intended subjects. For example, in Fig. 2 (a), a photo of <asset0> and <asset1>' becomes a photo of panda and bowl'.

Thus, two prompts are used: one is the prompt with placeholders, and the other is the prompt with class tokens replacing placeholders. We then extract attention maps from the T2I model associated with two prompts. To regularize the attention between subject placeholders and input images, we introduce *Inter-Cross-Attention (ICA)* loss. The ICA loss, denoted by $\mathcal{L}_{\text{ICA}}$, is defined as

$$\mathbb{E}_{z,s,t} \left[\sum_{i \in s} ||A_{t,i}(\gamma_\theta(c), z_t) - g(M_i \odot A_{t,i}(\gamma_\theta(\hat{c}), z_t))||_2^2 \right] \quad (3)$$

where c represents the sentence tokens with placeholders, $\hat{c}$ has placeholders replaced by class names, γ_θ is the text encoder, and i indexes the placeholders. The function $g(x) = x/\max(x)$ normalizes the attention map. To suppress activations beyond subject boundaries in the attention map generated by the prompt with class tokens, we element-wise multiply the attention map by the masks and then normalize it. The attention maps of prompts with class tokens are
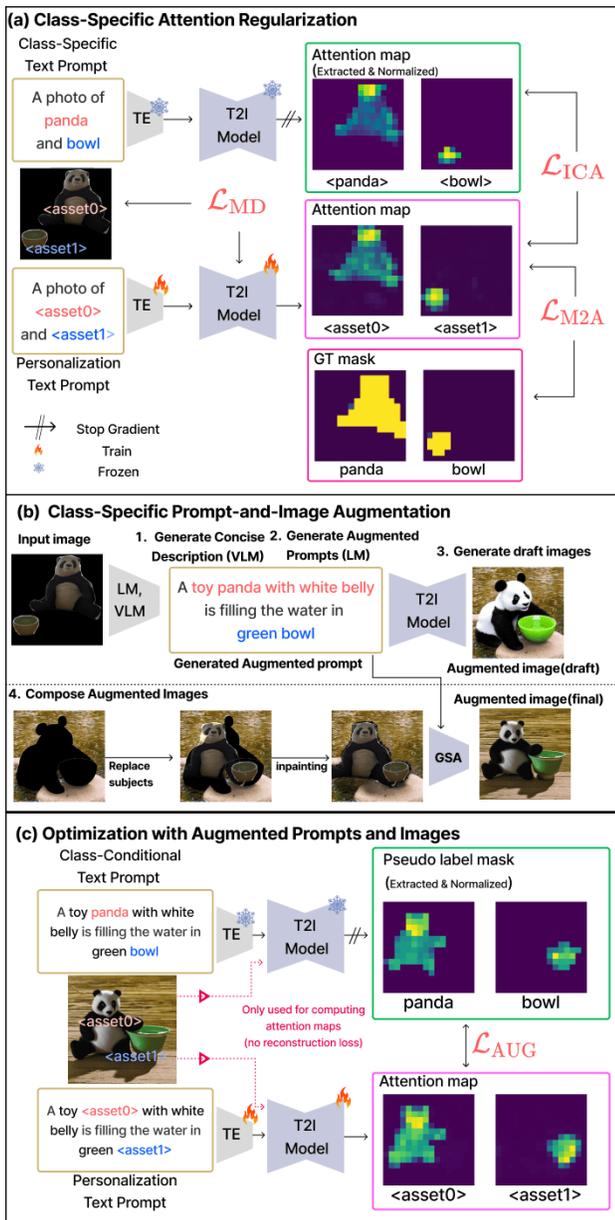
Class-Specific Prompt-and-Image Augmentation

Next, we focus on aligning the subject *actions* with class-specific priors. While the ICA loss mitigates overfitting to subject appearances, it remains limited in freely generating subject actions such as novel poses or interactions among subjects. In order to generate flexible variations in actions, we propose concept-based prompt-and-image augmentation. The technique utilizes the vast prior knowledge about object concepts contained within large-scale models, such as large (vision) language models, T2I models, etc.

Step 1: Generate Concise Description. We query a pre-trained vision-language model (VLM) about the input image to personalize. The VLM is expected to provide a detailed description of the subjects in a single sentence. However, it is important to use a specific prompt that limits the number of words to encourage concise responses. We obtain a concise description of a subject in the form of a *noun phrase*. Specifically, we input the subject image and the following prompt to the VLM: "`Tell me each subject in the photo less than 5 words using the noun phrase.`". In the example of Fig. 2(b), the generated noun phrases associated with `<asset0>` (resp. `<asset1>`), denoted by `<P1>` (resp. `<P2>`), is `toy panda with white belly` (resp. `green bowl`).

Step 2: Generate Augmentation Prompts. Based on concise descriptions of each subject from VLM, we generate multiple sentences on subject actions using a language model (LM). The goal is to generate prompts describing inter-subject interactions or individual subjects performing diverse actions. The input to LM can be neutral sentences such as "`Generate sentences of <P0> and <P1> interacting.`" where `<P0>` and `<P1>` are replaced with the noun phrases generated in **Step 1**. The text box in Fig. 2(b) shows an example of generated prompt. This is the *prompt augmentation*, and the augmented prompts will be used for training the generation model.

Step 3: Generate Draft Images. We input the augmented prompt to a T2I model and synthesize the associated images. In addition, pseudo-label masks for each subject are obtained by passing the rendered images through a generic, zero-shot segmentation model, e.g., Segment Anything Model (SAM) (Kirillov et al. 2023). The images generated in this step are based on general concepts (e.g., panda) and may not align well with the original input image for personalization. Thus, the generated image is used as a draft for composing augmented images in the next step.

Step 4: Compose Augmented Images. To better align

Figure 2: Overview of the DynASyn. (a) Concept-based Attention Regularization: the attention map derived from concept priors is used to regularize the attention map from token placeholder to prevent overfitting. (b) Concept-based Prompt-and-Image Augmentation: prompt-and-image pairs containing diverse action and poses of subjects are composed. (c) Optimization with Augmented Prompts and Images: the augmented data from (b) is used for our model to learn to generate novel actions and poses of the subject.

obtained by frozen models with stop-gradient: see Fig. 2(a). Such attention maps regularize the cross-attention between subject placeholders and input images, enabling the model to capture both subject identity and class-specific priors.
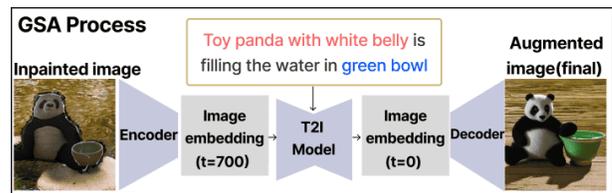
draft images with the original subject identity, we perform a series of operations as follows: see Fig. 2(b)-4. First, the subjects in draft images are replaced with the original version, ensuring spatial coherence. Then, any resulting gaps are filled using inpainting (Yu et al. 2023).

Next, inspired by SDEdit (Meng et al. 2021), we apply Guided SDE Augmentation (GSA), which is based on Stochastic Differential Equations (SDEs), to guide the T2I model with the augmented prompts. GSA starts with the inpainted image and performs 700 forward steps of diffusion. Then, it executes 700 backward steps using the T2I model conditioned on the augmented prompt from **Step 2** to obtain the final augmented image: see Fig. 3. GSA is specifically designed to align the augmented subject with the original through forward-backward SDE guided by augmented prompts (Fig. 3). This process ensures that the final augmented image is better aligned with the original subject compared to draft images, while still being capable of depicting various actions and poses through the guidance provided by the augmented prompts. The final augmented image is not a complete representation of personalization, but rather serves as an "interpolation" of diffusion sampling between the draft image and the inpainted image. For an example, compare the three images on the right of Fig. 2(b). In our implementation, we used GPT-4 (Achiam et al. 2023) as VLM and LM, Stable Diffusion v2.1 (Rombach et al. 2022) as the T2I model, and SAM (Kirillov et al. 2023) as the segmentation model.

Optimization with Augmented Prompts & Images

Using the augmented prompts and images, we optimize the placeholder embeddings, text encoder and T2I model. As previously, the cross-attention maps provided by the T2I model are utilized, where the attention maps between the augmented prompts and images are expected to capture various contextual cues and actions. We perform class-conditional attention regularization based on the augmented data as follows. Two prompts are used: one is the augmented prompt, and the other is the augmented prompt with the class token of the subject replaced by the placeholder token. Next, we extract attention maps from the T2I model using these sentences and augmented images, e.g., see Fig. 2 (c). If a reconstruction loss associated with input image is used, it can cause identity blending, because the model will be heavily influenced by individual pixel values from the augmented images. To maximize identity preservation while still allowing for action learning, we use only the inter-cross-attention loss without the reconstruction loss. The loss associated with the augmented data, denoted by $\mathcal{L}_{\text{AUG}}$, is defined as

$$\mathbb{E}_{z,s,t}\left[\sum_{i \in s} \|A_{t,i}(\gamma_\theta(c_a), z_t) - g(M_i \odot A_{t,i}(\gamma_\theta(\hat{c}_a), z_t))\|_2^2\right] \quad (4)$$

where c_a denote the text embeddings of the augmented prompt with placeholder, $\hat{c}_a$ denote the embeddings of the same prompt with placeholders replaced by class tokens, and M_i is the pseudo-label mask from the augmented image.

The final training objective is given by

$$\mathcal{L} = \lambda_{\text{MD}}\mathcal{L}_{\text{MD}} + \lambda_{\text{M2A}}\mathcal{L}_{\text{M2A}} + \lambda_{\text{ICA}}\mathcal{L}_{\text{ICA}} + \lambda_{\text{AUG}}\mathcal{L}_{\text{AUG}} \quad (5)$$

where λ_{MD}, λ_{M2A}, λ_{ICA}, and λ_{AUG} are hyperparameters for balancing the losses. By adjusting these hyperparameters, we strike a balance between identity regularization (e.g., $\mathcal{L}_{\text{MD}}$ and $\mathcal{L}_{\text{ICA}}$) and augmentation diversity (e.g., $\mathcal{L}_{\text{AUG}}$), ensuring that the model learns to preserve subject identity while generating diverse actions and interactions.

Experiments

Experimental Setup

Dataset. We train and evaluate on a total of 15 image-text pairing datasets, including 7 individual datasets from Break-a-Scene, 2 datasets from the COCO benchmark (Lin et al. 2014), 2 datasets from the Zhang et al (Zhang et al. 2024) and 4 additionally collected proprietary datasets. All datasets contain at least two distinct subjects. For the prompt-and-image augmentation, we generate 30 text prompts per image depicting new poses or inter-subject interactions using GPT-4 (Achiam et al. 2023).

General Settings. For all the methods, we utilized Stable Diffusion v2.1 (Rombach et al. 2022). We utilize Textual Inversion, Dreambooth, and Break-a-Scene as baselines for comparison. Since Textual Inversion and Dreambooth only allow single subject personalization, we use segmentation masks to restrict them to observing only the single subject per training step. Break-a-Scene enables multi-subject personalization, and thus we used it without modification. Detailed settings of hyperparmaeters are provided in Supplementary Materials.

Evaluation Metrics. We chose the metrics which gauge the similarity of generated images to the target subject and faithfulness to the conditioning text. We assessed the subject similarity via CLIP-I scores and text alignment through both CLIP-T scores and Image Reward (IR) scores (Xu et al. 2024). The IR score stems from a reward model trained on human judgments on text reflection and aesthetic quality for numerous text-image pairs.

Results. We conducted a comparison of the results for plain, action, and interaction by measuring text alignment and image similarity scores. Additionally, we performed a quantitative and qualitative comparison through user interviews obtained from 15 recruited volunteers. As shown in Table 1, DynASyn achieves the highest text alignment, as measured by CLIP-T and image reward scores in all of the plain, action, and interaction sentences. As expected, TI (Textual Inversion) (Gal et al. 2022) and DB (Dreambooth) (Ruiz et al. 2023a), tend to be ineffective for multi-subject or single image personalization due to poor text reflection. Despite being able to personalize a single image, Break-a-Scene (Avrahami et al. 2023) exhibits lower CLIP-T and image reward scores due to overfitting to appearances. In contrast, DynASyn resolves overfitting through prior-based personalization, yielding superior text alignment. DB attains the highest CLIP-I score since it is overfitted to the input, causing CLIP-I to favor output similarity, but at the cost of disregarded text conditions. Besides, DynASyn produces reasonable CLIP-I scores.

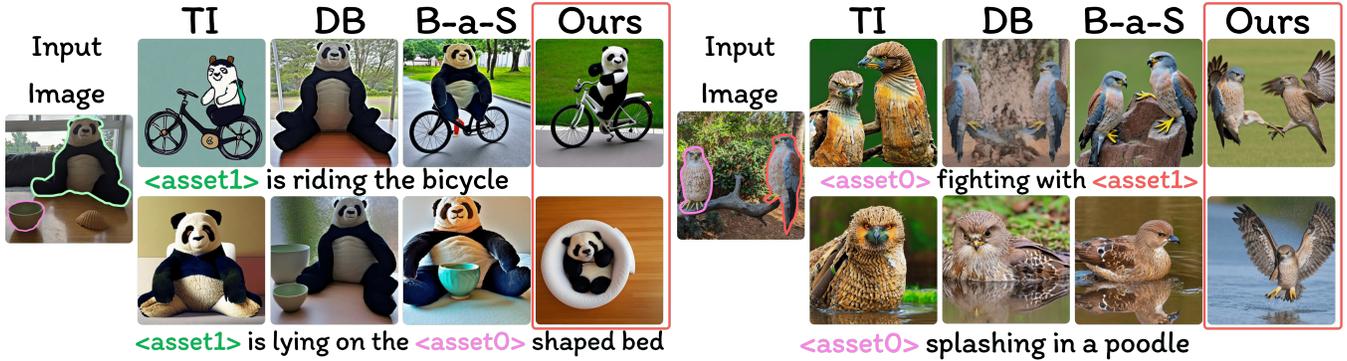This trend is also observed in the user interviews. DynASyn demonstrates superior performance, showing the

Figure 4: Qualitative comparisons with baseline methods. While baseline models often fail to align effectively with the provided text, DynASyn generates images that accurately reflect the textual input.

	Plain Sentence			Action Sentence			Interaction Sentence		
	IR ↑	CLIP-T ↑	CLIP-I ↑	IR ↑	CLIP-T ↑	CLIP-I ↑	IR ↑	CLIP-T ↑	CLIP-I ↑
TI (+mask)	0.45	0.285	0.71	0.41	0.252	0.688	0.208	0.225	0.659
DB (+mask)	0.713	0.306	**0.84**	0.509	0.274	**0.79**	0.459	0.247	**0.823**
B-a-S	_0.837_	_0.351_	0.774	_0.604_	_0.31_	0.732	0.58	_0.296_	0.679
DynASyn	**0.901**	**0.376**	_0.797_	**0.827**	**0.359**	_0.771_	**0.758**	**0.346**	_0.731_

Table 1: Quantitative comparisons with baseline methods. **Plain sentences** are those with no significant change in action. **Action sentences** describe a change in behavior for a single subject. **Interaction sentences** involve multiple subjects interacting. Prompt fidelity (Image Reward Score (IR), CLIP-T), Subject fidelity (CLIP-I)

	Plain Sentence			Action Sentence			Interaction Sentence		
	Overall ↑	Text ↑	Identity ↑	Overall ↑	Text ↑	Identity ↑	Overall ↑	Text ↑	Identity ↑
TI (+mask)	5.8	6.6	4.7	5.2	4.7	6.3	4.1	2.3	4.3
DB (+mask)	3.7	4.2	**8.1**	5.5	3.2	**8.6**	5.6	5.0	**7.6**
B-a-S	_6.2_	_7.5_	6.6	_6.0_	_6.5_	6.3	_6.2_	_5.5_	6.1
DynASyn	**8.2**	**8.5**	_7.4_	**7.5**	**7.4**	_7.5_	**7.7**	**7.7**	_7.3_

Table 2: Results of the user study. Identity measures how similar the subject in the generated image is to the original image, text evaluates how well the generated image reflects the textual input, and overall assesses the overall quality of the image.

highest overall and text alignment scores across all sentence types—Plain, Action, and Interaction—thereby validating its effectiveness. In contrast, TI and DB exhibit significantly lower overall and text alignment scores, with particularly poor performance on Action and Interaction sentences. While Dreambooth achieves the highest identity score, this is likely due to overfitting issues. Break-a-Scene shows lower text alignment and overall quality compared to DynASyn. Qualitative results in Fig. 4 show the comparison with the baseline methods[1], which confirms the afore-

---

[1]We attempted to make comparisons with recent works (Kumari et al. 2023) and (Zhang et al. 2024). A crucial step in those works involves creating a regularization image set similar to the input image. The step relied on the LAION dataset (Schuhmann et al. 2022) which however was taken down since December 2023 and was unavailable at the time of this writing. Thus, we inspected the generated samples in (Kumari et al. 2023) and (Zhang et al. 2024) instead of a direct comparison. We find that, while both of these works personalize well, the personalization prompts primarily involved changing styles or scenes, or inserting new elements,

MD	App	Act	IR ↑	CLIP-T ↑	CLIP-I ↑
✓			0.341	0.292	0.767
✓	✓		0.616	0.298	**0.786**
✓		✓	0.732	0.306	0.76
✓	✓	✓	**0.829**	**0.360**	0.767

Table 3: Ablation study

mentioned trends. DynASyn personalizes subjects without overfitting and reflects text well. TI struggles with identity preservation. DB is overfitted to the input image, barely modifying outputs based on text. Break-a-Scene reflects text more poorly although less overfitting than DB.

Ablation Study. We conducted an ablation study to validate the efficacy of DynASyn components, comparing four

---

but with few examples of altering poses or showing dynamic interactions with the surroundings. In contrast, our model is capable of generating a diverse range of actions by the subjects.

Figure 5: Visualization of personalized images. DynASyn generates a variety of images based on text when given a single input image. It can depict multiple subjects interacting dynamically or performing actions. Examples of additional generation tasks, such as re-contextualization or artistic stylization, can be found in Supplementary materials.
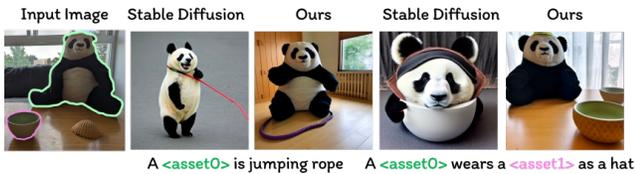

Figure 6: Limitation. In Stable Diffusion, We used class noun instead of placeholder `<asset>` for sampling.

model variants. The default model uses only masked diffusion ("MD" in Table 3), while we toggle the alignment of subject appearances (Step (a) in Fig. 2, "App") and actions (Step (b), (c), "Act"). As shown in Table 3, the CLIP-I scores show little overall difference across methods, which shows that overfitting was avoided while the identity is maintained. The CLIP-T scores, which measure the alignment of the text, also remain similar but improve slightly by adding the "App" component, which reduces the appearance overfitting. The inclusion of the "Act" component further enhances CLIP-T by learning various behaviors, with the full model achieving the highest CLIP-T and IR scores by balancing overfitting and diverse behavior learning.

Conclusions and Limitations

In this paper, we introduce DynASyn, a novel approach that leverages concept-specific regularization of cross-attention maps to effectively maintain identity representation while generating high-quality renditions across a wide array of prompt contexts. To mitigate overfitting, we have developed a prompt-image augmentation technique that creates diverse prompt-image pairs encompassing a variety of actions and interactions, thereby enhancing the model's robustness and generalization capabilities. Despite its promising performance, DynASyn presents certain limitations, as it is inherently dependent on prior knowledge and the performance of the underlying Text-to-Image (T2I) backbone. As illustrated in Fig. 6, models like Stable Diffusion struggle with challenging prompts, leading to a bottleneck where the accurate alignment of subject actions with concept priors is heavily reliant on the quality of generated images. To address these issues, it is necessary to explore additional fine-tuning strategies that could further refine the model's adaptability. Future work should focus on fine-tuning the T2I model using high-quality, expansive datasets, which will likely overcome the current limitations and facilitate more reliable and versatile image generation capabilities.

Acknowledgements

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Avrahami, O.; Aberman, K.; Fried, O.; Cohen-Or, D.; and Lischinski, D. 2023. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, 1–12.

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

Chen, M.; Laina, I.; and Vedaldi, A. 2024. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5343–5353.

Ding, M.; Zheng, W.; Hong, W.; and Tang, J. 2022. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35: 16890–16902.

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

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

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

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*.

Hua, M.; Liu, J.; Ding, F.; Liu, W.; Wu, J.; and He, Q. 2023. DreamTuner: Single Image is Enough for Subject-Driven Generation. *arXiv preprint arXiv:2312.13691*.

Jia, X.; Zhao, Y.; Chan, K. C.; 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 preprint arXiv:2304.02642*.

Kang, M.; Zhu, J.-Y.; Zhang, R.; Park, J.; Shechtman, E.; Paris, S.; and Park, T. 2023. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10124–10134.

Kim, C.; Lee, J.; Joung, S.; Kim, B.; and Baek, Y.-M. 2024. InstantFamily: Masked Attention for Zero-shot Multi-ID Image Generation. *arXiv preprint arXiv:2404.19427*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollar, P.; and Girshick, R. 2023. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4015–4026.

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, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22511–22521.

Li, Z.; Min, M. R.; Li, K.; and Xu, C. 2022. Stylet2i: Toward compositional and high-fidelity text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18197–18207.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Matsuda, H.; Togo, R.; Maeda, K.; Ogawa, T.; and Haseyama, M. 2024. Multi-Object Editing in Personalized Text-To-Image Diffusion Model Via Segmentation Guidance. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8140–8144. IEEE.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

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

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

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

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023a. 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.

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

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

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

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

Wang, Q.; Jia, X.; Li, X.; Li, T.; Ma, L.; Zhuge, Y.; and Lu, H. 2024a. StableIdentity: Inserting Anybody into Anywhere at First Sight. *arXiv preprint arXiv:2401.15975*.

Wang, Z.; Li, A.; Xie, E.; Zhu, L.; Guo, Y.; Dou, Q.; and Li, Z. 2024b. Customvideo: Customizing text-to-video generation with multiple subjects. *arXiv preprint arXiv:2401.09962*.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15943–15953.

Wei, Z.; Su, Q.; Qin, L.; and Wang, W. 2024. MM-Diff: High-Fidelity Image Personalization via Multi-Modal Condition Integration. *arXiv preprint arXiv:2403.15059*.

Wu, Z.; Yu, C.; Zhu, Z.; Wang, F.; and Bai, X. 2023. Singleinsert: Inserting new concepts from a single image into text-to-image models for flexible editing. *arXiv preprint arXiv:2310.08094*.

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

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2024. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3): 5.

Yu, T.; Feng, R.; Feng, R.; Liu, J.; Jin, X.; Zeng, W.; and Chen, Z. 2023. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*.

Zhang, Y.; Yang, M.; Zhou, Q.; and Wang, Z. 2024. Attention Calibration for Disentangled Text-to-Image Personalization. *arXiv preprint arXiv:2403.18551*.