CoCoIns: Consistent Subject Generation via Contrastive Instantiated Concepts

Anonymous authors
Paper under double-blind review

Figure 1. **Contrastive Concept Instantiation (CoCoIns)** is a generation framework achieving subject consistency across multiple individual creations without tuning or reference. Unlike prior work that requires customization tuning, adopts an additional encoder for reference, or generates images in batches, CoCoIns creates *instances of concepts* with a unique association that connects latent codes to subject instances. Given a latent code (o and x), CoCoIns converts it into a pseudo-word ([o] and [x]) that decides the appearance of a subject concept. By reusing the same code, users can consistently generate the same subject instances across multiple creations.

Abstract

While text-to-image generative models can synthesize diverse and faithful content, subject variation across multiple creations limits the application in long content generation. Existing approaches require time-consuming tuning, references for all subjects, or access to other creations. We introduce Contrastive Concept Instantiation (CoCoIns) to effectively synthesize consistent subjects across multiple independent creations. The framework consists of a generative model and a mapping network, which transforms input latent codes into pseudo-words associated with certain instances of concepts. Users can generate consistent subjects with the same latent codes. To construct such associations, we propose a contrastive learning approach that trains the network to differentiate the combination of prompts and latent codes. Extensive evaluations of human faces with a single subject show that CoCoIns performs comparably to existing methods while maintaining higher flexibility. We also demonstrate the potential of extending CoCoIns to multiple subjects and other object categories. The source code and models will be released.

1 Introduction

Text-to-image generation has made remarkable advances (Rombach et al., 2022; Saharia et al., 2022; Podell et al., 2023), opening up numerous downstream possibilities, including editing and style transfer. Among all applications, maintaining subject consistency has been a long-standing problem for long content creation, including storytelling (Li et al., 2019), comics (Wu et al., 2024a), or movie generation (Tulyakov et al., 2018; Polyak et al., 2025). These applications consist of sequences of images and clips, where consistent characters and objects facilitate recognizing subjects across moments and following the narratives.

While numerous approaches have been explored to ensure subject consistency, they are often labor-intensive or time-consuming. One straightforward approach is to gather all existing creations and manually swap generated subjects with reference subjects using face swapping (Nirkin et al., 2019; Bitouk et al., 2008). Another direction is to customize generators by optimizing virtual word tokens or finetuning model weights to represent reference subjects and produce new creations (Gal et al., 2023a; Ruiz et al., 2022). To reduce overhead for tuning generators, recent methods (Wei et al., 2023; Ye et al., 2023) incorporate additional encoders that convert references into representations. However, these methods still require users to prepare references for all subjects.

In contrast to addressing each creation individually, one can generate all creations in a batch, allowing samples within the batch to interact and achieve consistency (Tewel et al., 2024; Zhou et al., 2024; Liu et al., 2025). Specifically, the target prompts are merged into the same batch, and the latents of all samples are processed together, allowing subjects within the same batch to converge toward a similar appearance. Although promising results are achieved, these approaches require storing generated results to recreate the same subjects in the future.

We propose a generation framework that maintains subject consistency across individual creations without manual swapping, tuning, and reference preparation. Building such a framework presents numerous challenges – While we aim to generate a consistent subject appearance from a concept, preserving diversity among all instances of the concept remains important. Since the generator is already trained and exhibits high diversity and generalizability, we need to strike a good balance between *minimizing the variation* among the same subject instances across individual creations while *maintaining the diversity* between different instances. Additionally, collecting large-scale, high-quality data organized by subjects is challenging. Training the generator to synthesize annotated subjects in low-quality datasets directly could hamper both output quality and diversity.

To minimize variation among instances of the same subject while preserving diversity among instances, we introduce a latent space to model the distributions of instances for each concept. The proposed method is motivated by the common practices in natural and programming languages. If a user provides sufficient descriptions that encompass every intricate aspect of a concept, the generator may be able to consistently output the same appearance. Although covering comprehensive details is implausible with the limited vocabulary of human language, prior work on customizing generative models has shown the efficacy of pseudo-words (Gal et al., 2023a; Ruiz et al., 2022), which can convey essential information to represent particular subject instances. Our framework is built upon *instantiating concepts* (Anderson et al., 1976; Dershowitz, 1985) via pseudo-words. We associate codes in the latent space with specific concept instances in the output space as if we create instances identified by latent codes. These latent codes are embeddings sampled from the space, taken as input by the generation framework, and transformed into pseudo-words that guide the generator to synthesize specific instances.

To establish the association between input latent codes and output subject instances, we develop a lightweight mapping network that converts a latent code into a pseudo-word, which is then combined with a concept token to represent a specific instance of the concept. We then develop a contrastive learning strategy to train the mapping network. Instead of relying on subject annotations, the model learns to differentiate latent codes by comparing its own outputs generated from various combinations of prompts and latent codes. This self-supervision paradigm enables potential scalability and generalizability while avoiding the need to learn directly from limited data.

We refer to our generation framework as ***Contrastive Concept Instantiation (CoCoIns)***. As illustrated in Figure 1. Given a concept in a prompt indicating a subject (*e.g.* woman), the framework creates an instance of that concept by transforming a sampled latent code into a pseudo-word (*e.g.* [o] in the example prompt) that describes the concept. Each latent code and its transformed pseudo-word is uniquely tied to a specific instance and can be utilized for future creations. Different latent codes yield different instances, showcasing the preserved output diversity.

We conduct experiments on human images and perform systematic evaluation, including generating portrait photographs and free-form images. We achieve favorable subject consistency and prompt fidelity against

batch-generated approaches. We also demonstrate an early success in extending our approach to multi-subject and general concepts. The main contributions of this work are:

- To the best of our knowledge, we propose the first subject-consistent generation framework for multiple individual creations without tuning or encoding references.
- We develop a contrastive learning method that avoids learning from limited subject annotation and preserves output quality and diversity.
- We perform extensive evaluations and demonstrate favorable performance against approaches that require time-consuming tuning or batch generation.

2 Related Work

Subject-Driven Generation. One approach to subject consistency is based on subject-driven generation, which aims to generate customized topics according to user-provided input. By learning new tokens (Gal et al., 2023a; Voynov et al., 2023; Tewel et al., 2023) or model weights (Ruiz et al., 2022; Kumari et al., 2023; Han et al., 2023; Ruiz et al., 2024), pretrained generative models can be customized to produce outputs based on specific references. Textual Inversion (Gal et al., 2023a) learns virtual tokens that capture subject information inverted from reference images. DreamBooth (Ruiz et al., 2022) fine-tunes the parameters of pretrained models and learns unique identifiers that represent references. While subject consistency can be achieved by customizing each target concept with user-provided references, these methods are time-consuming as they require tuning for every subject.

To reduce the time and computational cost of tuning-based methods, another line of research incorporates additional encoders to obtain representations from reference images, which generative models take as conditions via augmented prompt embeddings (Wei et al., 2023; Shi et al., 2024; Xiao et al., 2024; Li et al., 2023; Wang et al., 2024a; He et al., 2024; Chen et al., 2023; Gal et al., 2023b; Avrahami et al., 2023), self-attention (Ding et al., 2024; Wang et al., 2024b), or cross-attention (Wei et al., 2023; Shi et al., 2024; Jia et al., 2023; Ye et al., 2023; Wang et al., 2024b; 2025). In addition, some approaches (Valevski et al., 2023; Wang et al., 2024d; Li et al., 2024; Peng et al., 2024; Papantoniou et al., 2024; Wu et al., 2024b; Yue et al., 2024) focus solely on specific domains, *e.g.* human faces, in exchange for adopting more powerful encoders dedicated to those domains, *e.g.* face recognition models (Deng et al., 2022). While encoders ease the tuning process, users still need to prepare references for all target concepts. Designing specific mechanisms to insert reference features into generative models is also necessary. In contrast, our approach operates in the text embedding space, offering the potential to apply to extensive text-based generative models.

Subject-Consistent Generation. While subject-driven generation produces new creations featuring the same subjects as references, another line of work focuses on generating a set of images with consistent subjects from a set of prompts. The Chosen One (Avrahami et al., 2024) iteratively selects a cluster of similar images and finetunes the model using that cluster. Consistory (Tewel et al., 2024), JeDi (Zeng et al., 2024), and StoryDiffusion (Zhou et al., 2024) utilize the self-attention features of all samples in a batch. However, these approaches are less flexible as they require access to other samples or features when performing new generations. 1Prompt1Story (Liu et al., 2025) consolidates all prompts into a single lengthy prompt in a specific format, where multiple context settings for a subject follow a single description of that subject. This format limits the expressiveness of the prompts. In contrast, we achieve subject consistency with greater flexibility by treating each generation individually while retaining the complete prompts.

Storytelling. Generating coherent stories (Li et al., 2019) requires maintaining character consistency over time. Some approaches (Rahman et al., 2023; Pan et al., 2024; Liu et al., 2024; Shen et al., 2025; Maharana et al., 2022) utilize cross-attention to access information from previous frames and prompts. Others introduce a memory bank (Maharana et al., 2021), perform auxiliary foreground segmentation (Song et al., 2020), or generate particular reference characters (Gong et al., 2023; Shen & Elhoseiny, 2023; Wang et al., 2024e). These methods often involve training on storytelling datasets and focus on generating complete stories or continuing them. Our approach emphasizes equipping existing generative models with the ability to maintain subject consistency.

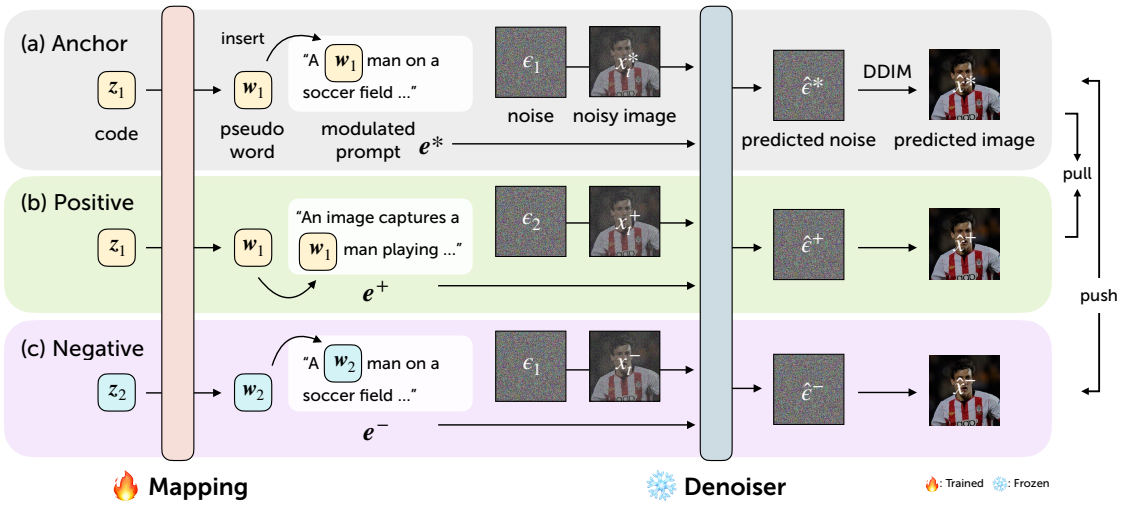

Figure 2. **Overview of Contrastive Concept Instantiation (CoCoIns).** We develop a contrastive learning approach to build associations between input latent codes and concept instances. For each training image, we generate two image descriptions and randomly sample two latent codes $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$. The mapping network first transforms the two latent codes into pseudo-words $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$. Then we collect a triplet of combinations of descriptions and latent codes. We build (a) an anchor sample with description embedding $\boldsymbol{e}^*$ modulated by inserting $\boldsymbol{w}_1$ before target concept token, (b) a positive sample $\boldsymbol{e}^+$ with a similar description embedding modulated with $\boldsymbol{w}_1$, along with (c) a negative sample $\boldsymbol{e}^-$ with the same prompt as the anchor but modulated with a different pseudo-word $\boldsymbol{w}_2$. The network is trained with a triplet loss to differentiate approximated images $\hat{\boldsymbol{x}}^*$, $\hat{\boldsymbol{x}}^+$, and $\hat{\boldsymbol{x}}^-$, from the denoiser prediction $\hat{\epsilon}^*$, $\hat{\epsilon}^+$, and $\hat{\epsilon}^-$.

StyleGAN. Our generation framework shares insights similar to StyleGAN (Karras et al., 2019; 2020). Both methods use a mapping network to transform input latents into an intermediate and more disentangled latent space. The space in StyleGAN enables better control over generated image attributes by modulating the generator through adaptive instance normalization (Huang & Belongie, 2017). In our framework, the intermediate latents operate in the same space as text embeddings, allowing for better manipulation of subject appearances through text conditions. CharacterFactory (Wang et al., 2024c) also learns a mapping network that transforms random noise into virtual tokens associated with human names defined in a celebrity dataset. These virtual tokens can produce consistent identities but lack control over semantics and attributes.

3 Methodology

Our goal is to maintain subject consistency across individual creations without time-consuming tuning or labor-intensive reference collections. We introduce ***Co**ntrastive **Co**ncept **In**stantiation* (CoCoIns), a generation framework that models concept instances in a latent space and uniquely associates latent codes in the space with output concept instances through contrastive learning. We introduce the base text-to-image model in Section 3.1. Then we describe the framework in Section 3.2 and the contrastive learning strategy in Section 3.3.

3.1 Text-to-Image Diffusion Models

We explore subject consistency in the context of text-to-image generation and base our approach on a latent diffusion model (Rombach et al., 2022; Podell et al., 2023). We utilize a pre-trained text-to-image model comprising an autoencoder (Kingma & Welling, 2014), a text encoder (Radford et al., 2021), and a denoiser. Given an image I and a prompt P, we obtain the latent image representation $\boldsymbol{x}$ and prompt embedding $\boldsymbol{e}$ via the autoencoder and text encoder, respectively. The denoiser ϵ_θ reverses the diffusion process:

$$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x} + \sqrt{1-\alpha_t}\epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \tag{1}$$

where $\alpha_{1:T} \in (0, 1]^T$ is a decreasing sequence, by predicting $\hat{\epsilon}$ from the noisy image $\boldsymbol{x}_t$, prompt $\boldsymbol{e}$, and timestep t:

$$\hat{\epsilon} = \epsilon_\theta(\boldsymbol{x}_t, \boldsymbol{e}, t). \tag{2}$$

3.2 Instantiating Concepts

To generate consistent instances of a concept, we model the distribution of instances in a latent space and associate latent codes in the space with concept instances in output images. As illustrated in Figure 2, a mapping network takes a latent code as input and produces a pseudo-word, which conveys necessary descriptive details to create a certain concept instance. The framework thus achieves subject consistency by generating the same subject with a fixed latent code over multiple creations.

The proposed mapping network transforms a latent code into a virtual word token, which is then inserted into the prompt embedding and guides the generation along with other words. Let $\boldsymbol{e} \in \mathbb{R}^{s \times d}$ denote the embedding of a prompt P obtained via dictionary lookup, where s is the sequence length. The concept token that a user wants to generate consistently (*e.g. man* in Figure 2) is at location i. Given a latent code $\boldsymbol{z} \in \mathbb{R}^c$, the mapping network $f : \mathbb{R}^c \to \mathbb{R}^d$ produces an pseudo-word embedding $\boldsymbol{w} \in \mathbb{R}^d$ that represents an instance:

$$\boldsymbol{w} = f(\boldsymbol{z}), \quad \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}). \tag{3}$$

Then we insert the output $\boldsymbol{w}$ into the prompt embedding $\boldsymbol{e}$ at the location i before the concept token and obtain the modulated prompt embedding $\hat{\boldsymbol{e}}$:

$$\hat{\boldsymbol{e}} = \texttt{insert}(\boldsymbol{e}, \boldsymbol{w}, i), \tag{4}$$

where `insert` denotes the insertion operation. The modulated prompt embedding $\hat{\boldsymbol{e}}$ is further encoded by the text encoder and serves as the text condition during generation.

3.3 Contrastive Association

We aim to establish unique associations between input latent codes and pseudo-words that represent visual instances in the output images. Thus, a latent code can be reused to generate the same concept instance.

A naïve way is to train the network f to synthesize subjects from a dataset with identity annotation, such as face recognition (Huang et al., 2008; Liu et al., 2015). However, we empirically find that training with common noise prediction (Ho et al., 2020) often compromises the generalizability and output quality of the generator, as these datasets are typically collected from data domains that are much narrower than the pre-trained data of the generator. The network learns to synthesize subjects from datasets, but it may overfit to the limited distributions. Thus, we develop a contrastive learning approach that does not require identity annotation, allowing us to train the mapping network in a self-supervised manner.

Constructing Triplets. We prepare multiple combinations of prompts and latent codes. As illustrated in Figure 2, the same prompts (*e.g.* "a man on a soccer field ...") are paired with different latent codes, and a similar prompt (*e.g.* "an image captures a man playing soccer ...") is coupled with the same latent code. The network is trained to generate pseudo-words inserted into prompts that synthesize certain instances.

Specifically, we prepare an image and a triplet of prompts for each training sample. The triplet consists of (a) an anchor prompt, (b) a positive prompt, and (c) a negative prompt. The anchor prompt is a caption that describes the image and is modulated by a latent code. The positive prompt is another description of the image modulated with the same latent code. The negative prompt has the same caption as the anchor prompt but is modulated with a different latent code. Formally, let $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ denote the embeddings of the two descriptions P_1 and P_2 of the image I. Here, $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ indicate two different latent codes. i and j are the locations of target concept tokens in $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$, respectively. We create an anchor prompt $\boldsymbol{e}^*$, a positive prompt $\boldsymbol{e}^+$, and a negative prompt $\boldsymbol{e}^-$ via

$$\begin{aligned} \boldsymbol{e}^* &= \texttt{insert}(\boldsymbol{e}_1, \boldsymbol{w}_1, i), \quad \boldsymbol{e}^+ = \texttt{insert}(\boldsymbol{e}_2, \boldsymbol{w}_1, j), \qquad \boldsymbol{w}_1 = f(\boldsymbol{z}_1), \\ \boldsymbol{e}^- &= \texttt{insert}(\boldsymbol{e}_1, \boldsymbol{w}_2, i), \qquad\qquad\qquad\qquad\quad \boldsymbol{w}_2 = f(\boldsymbol{z}_2). \end{aligned} \tag{5}$$

Then, we construct three noisy image latents to be paired with the three prompt embeddings. Since, in practice, a user may perform multiple generations with different initial noises, to maintain subject consistency between multiple generations, we obtain the noisy image latents by adding two different noises, ϵ_1 and ϵ_2, sampled from a normal distribution, to the image latent $\boldsymbol{x}$. We add the same noise to the anchor and negative samples to create a difficult situation:

$$\boldsymbol{x}_t^* = \sqrt{\alpha_t}\boldsymbol{x} + \sqrt{1-\alpha_t}\epsilon_1, \quad \boldsymbol{x}_t^+ = \sqrt{\alpha_t}\boldsymbol{x} + \sqrt{1-\alpha_t}\epsilon_2, \quad \boldsymbol{x}_t^- = \sqrt{\alpha_t}\boldsymbol{x} + \sqrt{1-\alpha_t}\epsilon_1, \tag{6}$$

where $\boldsymbol{x}_t^*$, $\boldsymbol{x}_t^+$, and $\boldsymbol{x}_t^-$ denotes the anchor, positive and negative noisy image latents, which are then paired with the modulated prompt embeddings $\boldsymbol{e}^*$, $\boldsymbol{e}^+$, and $\boldsymbol{e}^-$, respectively, to form the inputs to the denoiser.

Building Association. To encourage the generative model to synthesize subjects associated with pseudo-words, we apply a triplet loss (Schroff et al., 2015) to the denoiser outputs $\hat{\epsilon}^*$, $\hat{\epsilon}^+$, and $\hat{\epsilon}^-$ of the anchor, positive, and negative samples, pulling the anchor and positive samples closer while pushing away the negative sample. Since consistency is meaningful only in image latents instead of noise, we first acquire the predicted image latents $\hat{\boldsymbol{x}}^*$, $\hat{\boldsymbol{x}}^+$, and $\hat{\boldsymbol{x}}^-$ with DDIM (Song et al., 2021) approximation. Then, the distances between the three approximated latents are measured via

$$\mathcal{L}_{\text{con}} = \mathcal{L}_{\text{dis}}(\hat{\boldsymbol{x}}^*, \hat{\boldsymbol{x}}^+) + \lambda_{\text{neg}} \cdot \frac{1}{\mathcal{L}_{\text{dis}}(\hat{\boldsymbol{x}}^*, \hat{\boldsymbol{x}}^-)}, \tag{7}$$

where $\mathcal{L}_{\text{dis}}$ denotes a distance function. Note that since we empirically find that the common form of triplet with subtraction leads to less distinction between different input latent codes, the triplet loss is based on the reciprocal of the distance between the anchor and negative sample.

Subject Masks. Furthermore, since we only pursue subject consistency between images but not the similarity of entire images, we calculate the loss only in subject areas by applying masks to the output images. Subject masks can be obtained through an off-the-shelf referring segmentation model (Kirillov et al., 2023; Liu et al., 2023b; Ren et al., 2024), which annotates pixels corresponding to input words. Let m denote the mask with boolean values that cover the pixels of target concepts. We replace $\mathcal{L}_{\text{dis}}(\cdot,\cdot)$ with $\mathcal{L}_{\text{dis}}^m(\cdot,\cdot)$ in Eq. (7) to indicate the masked distance function with mask m, where $\mathcal{L}_{\text{dis}}^m(x,y) = \mathcal{L}_{\text{dis}}(m \cdot x, m \cdot y)$.

Background Preservation. With the aforementioned subject mask m, we negate the subject mask to acquire the background mask $\tilde{m} = 1 - m$. The background preservation loss is defined to minimize the distance between the backgrounds of the images generated with and without the virtual tokens:

$$\mathcal{L}_{\text{back}} = \mathcal{L}_{\text{dis}}^{\tilde{m}}(\hat{\boldsymbol{x}}^*, \hat{\boldsymbol{x}}_1) + \mathcal{L}_{\text{dis}}^{\tilde{m}}(\hat{\boldsymbol{x}}^+, \hat{\boldsymbol{x}}_2) + \mathcal{L}_{\text{dis}}^{\tilde{m}}(\hat{\boldsymbol{x}}^-, \hat{\boldsymbol{x}}_1). \tag{8}$$

Here $\hat{\boldsymbol{x}}_1$ and $\hat{\boldsymbol{x}}_2$ are DDIM approximation of the denoiser output $\epsilon_\theta(\boldsymbol{x}_t, \boldsymbol{e}_1, t)$ and $\epsilon_\theta(\boldsymbol{x}_t, \boldsymbol{e}_2, t)$, respectively. The final loss function consists of the contrastive and background preservation losses:

$$\mathcal{L} = \lambda_{\text{con}} \cdot \mathcal{L}_{\text{con}} + \lambda_{\text{back}} \cdot \mathcal{L}_{\text{back}}, \tag{9}$$

where λ_{con} and λ_{back} are weights balancing the two losses.

4 Experimental Results

4.1 Implementation Details

Architecture. We implement the mapping network f as an n-layered MLP and leverage Stable Diffusion XL (Podell et al., 2023) as the text-to-image diffusion model. We train only the mapping network f, with all the other model weights frozen.

Negative Distance Weight Schedule. We empirically find that the distance between anchor samples and negative samples is often too large at the beginning of training. The model tends to ignore the randomly initialized input latent codes and produces identical outputs. Therefore, we implement the weight of negative distance $\lambda_{\text{neg}} = \gamma(k/K)^\beta$ as an increasing function over training steps, where k and K are the current and total training steps, and γ as well as β are hyperparameters.

Figure 3. **Qualitative comparisons on *Portraits*** from (a) StoryDiffusion (Zhou et al., 2024), (b) Consistory (Tewel et al., 2024), (c) 1Prompt1Story (Liu et al., 2025), (d) DreamBooth (Ruiz et al., 2022), and (e) CoCoIns. The left four columns are generated with a man as the subject, and the right four with a woman. We achieve subject consistency without generating images in batches and through reference tuning.

Table 1. **Quantitative performance on *Portraits*.** We achieve comparable consistency and better diversity against the approaches that generate images in a batch.

	Sim↑	Div↑	CLIP↑
CelebA	0.590	0.992	0.299
Consistory	0.356	0.774	<u>0.218</u>
StoryDiffusion	**0.637**	0.577	0.217
1Prompt1Story	0.307	0.611	**0.228**
Ours	<u>0.600</u>	**0.799**	0.193

Table 2. **Quantitative performance on *Scenes*.** We achieve the best face similarity while maintaining similar subject diversity and prompt fidelity against other methods. In addition, we generate images individually, enabling high flexibility in future creations.

	Sim↑	Div↑	CLIP↑	DS↑
Consistory	0.098	**0.883**	**0.297**	0.383
StoryDiffusion	<u>0.159</u>	0.814	<u>0.290</u>	**0.407**
Ours	**0.256**	<u>0.847</u>	<u>0.290</u>	<u>0.388</u>

4.2 Experiment Setups

We conduct comprehensive experiments on single-subject human faces with our approach, followed by multiple subjects and other object categories. More details on data collection can be found in Appendix D.

Training. We train the token modulator using the CelebA dataset (Liu et al., 2015), which comprises 20K images and 10K identities. We generate prompts with the captioning model LLaVA-Next (Liu et al., 2023a) and masks with the zero-shot referring segmentation model Grounded SAM 2 (Kirillov et al., 2023; Liu et al., 2023b; Ren et al., 2024).

Evaluation. We design two prompt sets for experiments for comprehensive evaluations

- *Portraits*: This set evaluates face similarity with clear front faces. It contains 1K sentences composed of the template "A `[subject]` is looking at the camera.", where `[subject]` is one of `{man`, `woman`, `boy`, `girl`, `person}`. Each subject contains 200 sentences, resulting in a total of 1K samples
- *Scenes*: This set represents real-world performance with free-form prompts, where face poses and angles vary. It comprises 1K sentences generated by a Large Language Model (LLM) (Ouyang et al., 2022). We prompt the LLM to generate sentences of the same five subjects doing something in diverse situations, including four settings: daily lives, professional environments, cultural or recreational occasions, and outdoor activities, each with 50 samples.

We measure subject similarity and diversity over *Portraits* and *Scenes*. For faces, we estimate face similarity (Sim) and diversity (Div) of cropped and aligned images. Face similarity is the pairwise cosine similarity of ArcFace (Deng et al., 2022) embeddings between images of the same identities. To estimate diversity, we first average the face embeddings on the same identities. We then calculate the pairwise cosine similarity between the averaged embeddings of all identities.

Since *Scenes* considers images in real-world applications where faces are not always clear and large, we calculate DreamSim (Fu et al., 2023) (DS), a learned perceptual distance aligned with human preference, for subject similarity. We also measure prompt fidelity for both sets using CLIP, which is the cosine similarity between the projected embeddings of the CLIP text and image encoders.

Evaluated Methods. We evaluate our method against tuning-free subject-consistent generation and tuning-based customization. Tuning-free schemes include Consistory (Tewel et al., 2024), StoryDiffuion (Zhou et al., 2024), and 1Prompt1Story (Liu et al., 2025). Since tuning-based methods require training for all subjects, which incurs heavy computational costs, we use DreamBooth (Ruiz et al., 2022) as an example and present only quantitative results.

4.3 Empirical Results

Table 1 shows the quantitative performances of *Portraits*, and Figure 3 displays two subjects, a man and a woman, each with four images, generated by all approaches. We measure the similarity (Sim) and diversity (Div) of the training dataset CelebA for reference. Although StoryDiffusion exhibits the highest similarity, even surpassing CelebA, its diversity remains low. Subjects generated from different batches with different initial noise converge toward a similar appearance. Our approach achieves comparable similarity to StoryDiffusion and CelebA while generating diverse subjects.

We demonstrate the examples of *Scenes* in Figure 4 and present the quantitative performance in Table 2. We compare our approach against Consistory and StoryDiffusion because 1Prompt1Story does not support free-form prompts. It operates with a particular prompt structure where a subject description is followed by multiple context descriptions. Our model performs favorably in terms of face similarity while also performing comparably in terms of diversity and fidelity.

4.4 Ablation Study

Consistency Loss. We analyze the efficacy of triplet loss in maintaining consistency. First, we calculate face similarity and diversity on *Portraits* and prompt fidelity on *Scenes*. Table 3 shows the performance of applying the loss of only positive distance (Pos), positive and negative distance (Pos + Neg), using the common form of triplet loss with negative distance subtraction (subtract), and setting the weight of negative distance as a increasing sequence over training iterations. The four settings yield similar prompt fidelity; however, implementing negative distance, the reciprocal form, and the weighing schedule enhances performance.

Background Loss. We also evaluate the effect of removing background preservation loss under the same test sets. As shown in Table 4, only consistency loss results in unsatisfactory face similarity and fidelity. While training with masked distances (+ Mask) without preserving backgrounds achieves slightly higher face similarity, it attains low diversity and fidelity. In this setting, although the loss is only calculated within masked regions, the model modifies non-masked backgrounds, possibly due to the self-attention mechanism,

Figure 4. **Qualitative comparisons on *Scenes*** from (a) StoryDiffusion Zhou et al. (2024), (b) Consistory Tewel et al. (2024), (c) DreamBooth Ruiz et al. (2022), and (d) CoCoIns. 1Prompt1Story Liu et al. (2025) is absent here because it needs a specific prompt format with unified subject descriptions. The left and right four columns are two different subjects in diverse contexts. The prompts can be found in Appendix D.

Table 3. **Performance of ablating consistency loss.** We train the network with the distance from positive sample (Pos) and the reciprocal of negative sample (Neg). Instead of the common triplet loss with the subtraction of negative distance (subtract), we minimize its reciprocal and apply a weighing schedule increasing along training iterations (Schedule). The final setting achieves the best face similarity and diversity with similar prompt fidelity.

	Sim↑	Div↑	CLIP↑
Pos	0.394	0.500	0.293
Pos + Neg	0.492	0.750	0.290
Pos + Neg (subtract)	0.380	0.444	**0.294**
Pos + Neg + Schedule	**0.600**	**0.799**	0.290

Table 4. **Performance of ablating background Loss**. In addition to the triplet loss for consistency (Consistency Loss), we adopt a segmentation mask (Mask) to control the variation area and a background preservation loss (Background) to make backgrounds closed to the original predictions, which significantly improves face diversity and prompt fidelity with similar face similarity.

	Sim↑	Div↑	CLIP↑
Consistency Loss	0.444	0.395	0.138
+ Mask	**0.619**	0.352	0.128
+ Mask + Background	0.600	**0.799**	**0.290**

which allows information to interact globally. Applying background (+ Background) preservation significantly improves diversity and fidelity.

Prompt and Noise Combinations. We compare the strategy for constructing training triplets. We evaluate the performance of using the same two prompts or creating noisy latent images with the same noise for anchor and positive samples. Table 5 shows that using different prompts and noise achieves the best similarity and almost the same prompt fidelity.

Training with Subject Annotations. We show the results of directly training the mapping network as a noise prediction problem with CelebA. Figure 5 shows that the model learns to maintain subject consistency, but the output quality is also affected by limited, low-quality data.

Table 5. **Performance of prompts and noise combinations of constructing training triplets.** Compared to adopting the same prompts (=) or noise for the anchor and positive samples, using two different ($\neq$) prompts and noise yields the best face similarity and diversity with similar prompt fidelity.

Prompts	Noise	Sim↑	Div↑	CLIP↑
=	$\neq$	0.548	0.686	0.290
$\neq$	=	0.306	0.772	**0.292**
$\neq$	$\neq$	**0.600**	**0.799**	0.290

Figure 6. **Results of general concepts consistency.** Our approach makes no assumptions about object categories. It can be potentially applied to other concepts such as cats, dogs, and cars.

Figure 5. **Results of training the mapping network as a noise prediction problem.** The network overfits the dataset and generates images of low quality.

Figure 7. **Results of multi-subject consistency.** Given two different latent codes, the model trained with single-subject images can maintain consistency for multiple subjects.

4.5 Extensions

General Concepts. Since our approach does not impose any constraints on subject classes, we also demonstrate that it can be applied to general concepts. We train the model with animal (Choi et al., 2020) and cars (Yu et al., 2015) images. The examples in Figure 6 show that the model can potentially be applied to more concept categories besides humans.

Multi-Subject Consistency. In addition to analyzing single subjects, we demonstrate our potential to support consistency between images with multiple subjects. Figure 7 contains two sets of examples of a man and a woman in different settings. We use two different input codes for two subjects and generate the images with the model trained on single-subject data. While the model has never seen two faces in an image, it can identify face areas and maintain consistency for multiple subjects. Despite some entanglements and influence between subjects, the results demonstrate the potential to extend the model to multi-subject scenarios.

5 Conclusion

In this work, we propose Contrastive Concept Instantiation (CoCoIns), the first approach to achieve subject consistency without the need for time-consuming tuning and labor-intensive reference collection. Our key idea is to model concept instances in a latent space and train a mapping network with contrastive learning to associate latent codes in the space with output concept instances. We demonstrate its efficacy on single-subject human faces and extend it to multi-subject and general concepts. We believe this work establishes a foundation for ultimately controllable content creation.

References

Richard C. Anderson, James W. Pichert, Ernest T. Goetz, Diane L. Schallert, Kathleen V. Stevens, and Stanley R. Trollip. Instantiation of general terms. *Journal of Verbal Learning and Verbal Behavior*, 1976. 2

Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia*, 2023. 3

Omri Avrahami, Amir Hertz, Yael Vinker, Moab Arar, Shlomi Fruchter, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. The chosen one: Consistent characters in text-to-image diffusion models. In *SIGGRAPH*, 2024. 3

Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, and Shree K. Nayar. Face swapping: automatically replacing faces in photographs. In *SIGGRAPH*, 2008. 2

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. In *NeurIPS*, 2023. 3

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 10

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE TPAMI*, 2022. 3, 8

Nachum Dershowitz. Program abstraction and instantiation. *ACM Trans. Program. Lang. Syst.*, 1985. 2

Ganggui Ding, Canyu Zhao, Wen Wang, Zhen Yang, Zide Liu, Hao Chen, and Chunhua Shen. Freecustom: Tuning-free customized image generation for multi-concept composition. In *CVPR*, 2024. 3

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *NeurIPS*, 2023. 8, 17

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023a. 2, 3

Rinon Gal, Moab Arar, Yuval Atzmon, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM TOG*, 2023b. 3

Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Yingqing He, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, and Yujiu Yang. Interactive story visualization with multiple characters. In *SIGGRAPH Asia*, 2023. 3

Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *ICCV*, 2023. 3

Zecheng He, Bo Sun, Felix Juefei-Xu, Haoyu Ma, Ankit Ramchandani, Vincent Cheung, Siddharth Shah, Anmol Kalia, Harihar Subramanyam, Alireza Zareian, Li Chen, Ankit Jain, Ning Zhang, Peizhao Zhang, Roshan Sumbaly, Peter Vajda, and Animesh Sinha. Imagine yourself: Tuning-free personalized image generation. *arXiv:2409.13346*, 2024. 3

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. 22

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 5

Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on Faces in Real-Life Images: Detection, Alignment, and Recognition*, 2008. 5

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 4

Xuhui Jia, Yang Zhao, Kelvin C. K. 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:2304.02642*, 2023. 3

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 4

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 4

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 4

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *CVPR*, 2023. 6, 7

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 3

Dongxu Li, Junnan Li, and Steven C. H. Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv:2305.14720*, 2023. 3

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *CVPR*, 2019. 1, 3

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *CVPR*, 2024. 3

Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. Intelligent grimm - open-ended visual storytelling via latent diffusion models. In *CVPR*, June 2024. 3

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023a. 7, 19

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv:2303.05499*, 2023b. 6, 7

Tao Liu, Kai Wang, Senmao Li, Joost van de Weijer, Fahad Shahbaz Khan, Shiqi Yang, Yaxing Wang, Jian Yang, and Ming-Ming Cheng. One-prompt-one-story: Free-lunch consistent text-to-image generation using a single prompt. In *ICLR*, 2025. 2, 3, 7, 8, 9

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 5, 7, 22

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Improving generation and evaluation of visual stories via semantic consistency. In *NAACL*, 2021. 3

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *ECCV*, 2022. 3

Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *ICCV*, 2019. 2

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022. 8

Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhu Chen. Synthesizing coherent story with auto-regressive latent diffusion models. In *WACV*, 2024. 3

Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2face: A foundation model for id-consistent human faces. In *ECCV*, 2024. 3

Xu Peng, Junwei Zhu, Boyuan Jiang, Ying Tai, Donghao Luo, Jiangning Zhang, Wei Lin, Taisong Jin, Chengjie Wang, and Rongrong Ji. Portraitbooth: A versatile portrait model for fast identity-preserved personalization. In *CVPR*, 2024. 3

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv:2307.01952*, 2023. 1, 4, 6

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models. *arXiv:2410.13720*, 2025. 1

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv:2103.00020*, 2021. 4

Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. Make-a-story: Visual memory conditioned consistent story generation. In *CVPR*, 2023. 3

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv:2401.14159*, 2024. 6, 7, 22

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 4

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 *CVPR*, 2022. 2, 3, 7, 8, 9

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:2307.06949*, 2024. 3

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv:2205.11487*, 2022. 1

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 6

Fei Shen, Hu Ye, Sibo Liu, Jun Zhang, Cong Wang, Xiao Han, and Wei Yang. Boosting consistency in story visualization with rich-contextual conditional diffusion models. In *AAAI*, 2025. 3

Xiaoqian Shen and Mohamed Elhoseiny. Storygpt-v: Large language models as consistent story visualizers. *arXiv:2312.02252*, 2023. 3

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *CVPR*, 2024. 3

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 6

Yun-Zhu Song, Zhi Rui Tam, Hung-Jen Chen, Huiao-Han Lu, and Hong-Han Shuai. Character-preserving coherent story visualization. In *ECCV*, 2020. 3

Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *SIGGRAPH*, 2023. 3

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. In *SIGGRAPH*, 2024. 2, 3, 7, 8, 9

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *CVPR*, 2018. 1

Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. Face0: Instantaneously conditioning a text-to-image model on a face. In *SIGGRAPH Asia*, 2023. 3

Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. P+: Extended textual conditioning in text-to-image generation. *arXiv:2303.09522*, 2023. 3

Jiahao Wang, Caixia Yan, Haonan Lin, Weizhan Zhang, Mengmeng Wang, Tieliang Gong, Guang Dai, and Hao Sun. Oneactor: Consistent subject generation via cluster-conditioned guidance. In *NeurIPS*, 2024a. 3

Kuan-Chieh Wang, Daniil Ostashev, Yuwei Fang, Sergey Tulyakov, and Kfir Aberman. Moa: Mixture-of-attention for subject-context disentanglement in personalized image generation. In *SIGGRAPH Asia*, 2024b. 3

Qinghe Wang, Baolu Li, Xiaomin Li, Bing Cao, Liqian Ma, Huchuan Lu, and Xu Jia. Characterfactory: Sampling consistent characters with gans for diffusion models. *arXiv:2404.15677*, 2024c. 4

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv:2401.07519*, 2024d. 3

Wen Wang, Canyu Zhao, Hao Chen, Zhekai Chen, Kecheng Zheng, and Chunhua Shen. Autostory: Generating diverse storytelling images with minimal human effort. *IJCV*, 2024e. 3

Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. MS-diffusion: Multi-subject zero-shot image personalization with layout guidance. In *ICLR*, 2025. 3

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. In *ICCV*, 2023. 2, 3

Jianzong Wu, Chao Tang, Jingbo Wang, Yanhong Zeng, Xiangtai Li, and Yunhai Tong. Diffsensei: Bridging multi-modal llms and diffusion models for customized manga generation. *arXiv:2412.07589*, 2024a. 1

Yi Wu, Ziqiang Li, Heliang Zheng, Chaoyue Wang, and Bin Li. Infinite-id: Identity-preserved personalization via id-semantics decoupling paradigm. In *ECCV*, 2024b. 3

Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *IJCV*, 2024. 3

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv:2308.06721*, 2023. 2, 3

Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 10

Dongxu Yue, Maomao Li, Yunfei Liu, Qin Guo, Ailing Zeng, Tianyu Yang, and Yu Li. Addme: Zero-shot group-photo synthesis by inserting people into scenes. In *ECCV*, 2024. 3

Yu Zeng, Vishal M. Patel, Haochen Wang, Xun Huang, Ting-Chun Wang, Ming-Yu Liu, and Yogesh Balaji. Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation. In *CVPR*, 2024. 3

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. In *NeurIPS*, 2024. 2, 3, 7, 8, 9

A Comparison of Settings and Implementation

We provide a comparison of settings and implementation between our approach and prior work in Table 6. As discussed in Section 2, prior approaches can be categorized into subject-driven generation and subject-consistent generation.

Subject-driven approaches personalize a generative model by tuning parameters on reference images or incorporating additional pretrained encoders. These approaches are either time-consuming due to subject-specific tuning or require integration of general encoders such as DINO or domain-specific encoders like face recognition models.

Subject-consistent approaches modify the prompt or attention mechanisms of the base generator, avoiding extra modules, reference images, or additional training. Their main limitation is the requirement for generating images in batches or reliance on stored features.

Our method is more lightweight and flexible. It requires only an MLP trained once and supports individual inference.

Table 6. Comparison of settings and implementation between our approach and prior work.

Approach	Extra Modules	Reference	Training	Inference Constraints
Tuning-based Personalization	None	Yes	Subject Tuning	None
Encoder-based Personalization	Pretrained Encoders	Yes	Once	None
Subject-Consistent Generation	None	No	None	Batch
CoCoIns (Ours)	Lightweight MLP	No	Once	None

B Additional Implementation Details

Computational Resources. The experiments are conducted on AMD EPYC 9354 CPU and four NVIDIA A6000 GPUs. Each training round takes around eight hours.

Hyperparameters. We list the hyperparameters in Table 7. They are decided by grid search. The distance function $\mathcal{L}_{\text{dis}}$ is the Mean Squared Error.

Latent Code Sampling. Latent codes are randomly sampled from Gaussian Distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$, stated in Equation (3). In each training triplet, the anchor and positive samples share the same latent code, and the negative sample is paired with another randomly sampled code, as detailed in Equation (5).

Table 7. Hyperparameters.

Hyperparameters	Value
c	256
λ_{cons}	1
λ_{back}	30
γ	0.00001
β	2
n	8
K	5000
Batch Size	128
Learning Rate	0.0001
Learning Rate Decay	Cosine
Learning Rate Warmup	500
Optimizer	Adam
Weight Decay	0.2

Table 8. FID score between each approach and its base models.

	FID ↓
Consistory	13.3
StoryDiffusion	15.2
Ours	9.2

Table 9. Comparison of performance using MSE and DreamSim loss as the distance function.

Distance	Sim ↑	Div ↑	CLIP ↑
MSE	0.600	0.799	0.193
DreamSim	0.314	0.724	0.291

Table 10. Ablation study on weighting background loss.

λ_{back}	Sim ↑	Div ↑	CLIP ↑
10	0.637	0.603	0.265
30	0.600	0.799	0.290
50	0.516	0.707	0.292

Table 11. Ablation study on weighting negative distances.

γ	Sim ↑	Div ↑	CLIP ↑
10^{-4}	0.666	0.766	0.288
10^{-5}	0.600	0.799	0.290
10^{-6}	0.445	0.635	0.293

Table 12. Ablation study on schedules of negative distance weighting.

β	Sim ↑	Div ↑	CLIP ↑
1	0.528	0.735	0.291
2	0.600	0.799	0.290
3	0.520	0.710	0.291

C Additional Experimental Results

C.1 Additional Comparison

Image Quality. We evaluate the quality of the generated images using the Fréchet Inception Distance (FID) score. Since various approaches are based on different models, we compute the FID scores by comparing the images produced by each approach with those generated by their respective base models. Consistory and we use SDXL; StoryDiffusion utilizes RealVisXL. Given that the FID score is sensitive to sample size, we duplicate the dataset of *Scenes* ten times to create a total of 10K prompts. We then generate 10K images using both the compared approaches and their base models. As shown in Table 8, our generated images are more closely aligned with the distribution of the base model.

Distance Function. We investigate the effect of different distance functions for measuring subject similarity during training. In our training strategy, the model is trained to generate similar appearances from the same image corrupted by two different noises, using two similar prompts and the same latent code. When reconstruction is nearly perfect, a simple pixel-wise metric such as Mean Squared Error (MSE) is sufficient. However, in more realistic scenarios where outputs are imperfect, we may need a subject-aware similarity measure. One option is to use a subject encoder; however, existing encoders are typically designed for narrow domains (e.g., face recognition) and often rely on non-differentiable operations, such as cropping and landmark alignment, making them unsuitable for end-to-end training. An alternative is a learned perceptual loss, such as DreamSim (Fu et al., 2023), which is trained to align with human perception of similarity. However, its notion of similarity may reflect factors like layout and color rather than subject identity.

Therefore, we primarily use MSE as our distance function, but also include a comparison against DreamSim. To apply DreamSim as a perceptual similarity metric, we decode the DDIM-approximated image latents using the Autoencoder and compute the cosine similarity between their DreamSim embeddings. We use the DINOv2 checkpoint for DreamSim. As shown in Table 9, compared to MSE, DreamSim loss struggles to generate consistent subjects.

C.2 Additional Ablation Study

We examine the impact of hyperparameter choices on the performance, specifically (1) the balance between the consistency and background loss and (2) the weighting scheduling of negative distances.

Table 10 shows the comparison between different background loss weightings, denoted by λ_{back} in Equation (9). Increasing the importance of background loss improves prompt fidelity at the cost of lower face similarity and diversity. We choose the λ_{back} that balances these two factors.

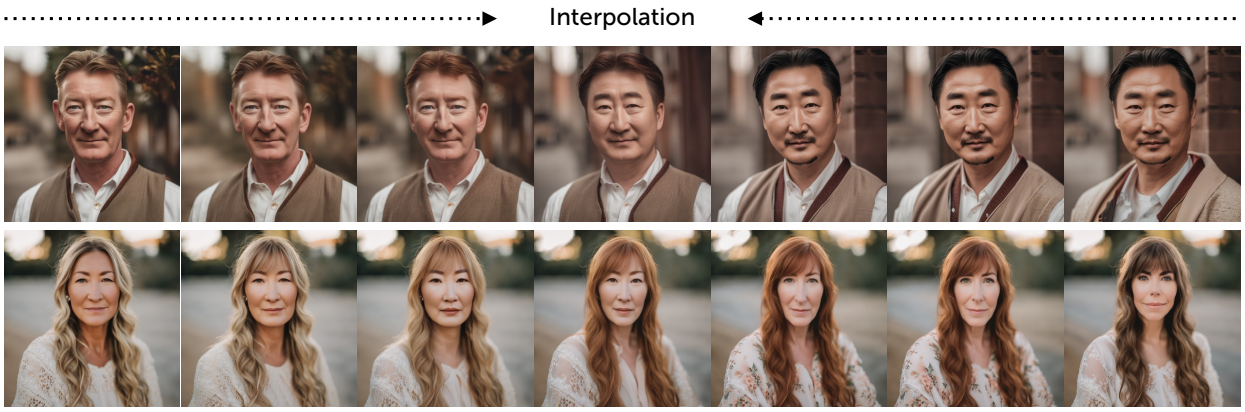

Figure 8. **Results generated with interpolations of two latent codes.** The leftmost and rightmost images are generated by two randomly sampled codes. The intermediate images are the results of the interpolations of the two codes, demonstrating the gradual transition between the two faces.

Figure 9. **Results generated with neighbors of a pseudo-word.** We generate 100 images with randomly sampled latent codes and a fixed initial noise. Given a randomly selected code and its corresponding image, we find the neighbors of the selected code by sorting the cosine similarity between the pseudo-words transformed from the code and the others. The result shows that the pseudo-words of the closer neighbors (*i.e.* high cosine similarity) produces more similar faces.

Additionally, as discussed in Section 4.1, the weighting of negative distances is implemented as an increasing schedule parameterized by $\gamma(k/K)^{\beta}$, where k is the current training step, and K is the number of total training steps. γ and β are hyperparameters. We also examine the effect of varying these two hyperparameters.

Table 11 presents the results for different values of γ, and Table 12 provides comparisons between different values of β. Similar to our observations in the previous study on the consistency and background loss, some sets of hyperparameters lead to higher face similarity but lower diversity or prompt fidelity. We choose the hyperparameter set that balances all of these important metrics.

C.3 Analysis of the Latent Space

We analyze the latent space that models the instance distributions of concepts. To help understand the relationships between latent codes in the space and the information captured by pseudo-words, we demonstrate the results generated with interpolations of latent codes and neighbors of pseudo-words.

Interpolation of Latent Codes. We reveal the relationships between latent codes by visualizing their interpolations. Given two randomly sampled latent codes, we generate images through their gradual interpolation and a fixed initial noise for the generation process. As illustrated in the two examples in Figure 8, the leftmost and rightmost images correspond to the two original latent codes, while the intermediate images

Figure 10. Eight sets of examples from *Portraits*. The subjects for each row are woman, man, girl, and boy.

depict the transition between them. The use of fixed initial noise results in a similar background appearance across the interpolated images.

Neighbors of Pseudo-Words. We also illustrate the information captured by a pseudo-word by retrieving its neighbors. Specifically, we randomly sample 100 latent codes and generate images from them. We then select one code and its corresponding image, and compute the cosine similarity between the pseudo-word derived from this code and those from the remaining codes. The other images are sorted based on this similarity. As shown in Figure 9, the leftmost image is generated from the selected latent code, while the images to its right are arranged in descending order of pseudo-word similarity. The results indicate that closer neighbors, *i.e.*, pseudo-words with higher similarity, tend to generate more visually similar faces.

C.4 Additional Generated Samples

We provide additional samples from *Portrait* in Figure 10 and *Scenes* in Figure 11. Each row contains two subjects with four examples. The subjects for each row are woman, man, girl, and boy.

More multi-subject examples are in Figure 12. We also include the comparison with two previous approaches, Consistory and StoryDiffusion. Our examples demonstrate high image quality and consistency. The prompts in Figure 7 and Figure 12 are both "a man and a woman sitting on the sofa", "a man and a woman taking the bus", "a man and a woman walking on the street", and "a man and a woman dancing on the stage".

More examples in Figure 13 demonstrate consistency for general concepts. The prompts in Figure 6 and Figure 13 are "a photo of a cat", "dog" or "car".

D Data Collection and Generation

Identifying Subjects in Descriptions. We generate descriptions for training images with a multi-modal captioner (Liu et al., 2023a). Then, we prompt an LLM to identify the words that are most likely to be the subjects in the sentences. The prompt is as follows:

> These are two captions of an image. Tag the words related to the subjects of the captions. Return the captions in a JSON with keys "caption1" and "caption2".

Figure 11. Eight sets of examples from *Scenes*. The subjects for each row are woman, man, girl, and boy.

Figure 12. More examples and comparison with previous work of multi-subject consistency.

1. Identify words related to a person: Look for specific nouns or noun phrases that refer to a person. Exclude pronouns (e.g., “he”, “she”, “they”, “it”) and collective nouns (e.g., “people”) from tagging.
2. Tag these words: Surround each identified word or noun phrase with `<subj>` and `</subj>` tags. Use `<subj1>`, `<subj2>`, etc., for different items. Apply the same index number to all references to the same item.
3. If no relevant words are found: Return the original captions without any changes.

Figure 13. Two sets of examples demonstrating consistency for general concepts.

4. Handling ambiguity: If a word has ambiguous references or unclear roles, do not tag it.
5. Pronouns: Do not tag pronouns.

Then, we implement a tag parser to provide subject locations during training. In inference, users can provide the locations to indicate target concepts that need to be consistent.

Generating Diverse Scenes for Evaluation. One of the test datasets is collected to represent diverse, real-world contexts. Taking "man" as an example, we prompt an LLM to generate these sentences using the following instruction:

> Generate fifty scene descriptions of a man in daily life.
>
> 1. Focus on the subject
>
> - The subject should always be a "man".
> - Provide descriptions of diverse scenes that feature a specific subject performing an action in a certain place.
>
> 2. Make actions and locations diverse
>
> - Always use a verb and location that has not appeared in previous sentences.
> - The scenes should be related to everyday life, such as cooking, driving, walking, reading, etc.
>
> 3. Portrait Details:
>
> - The descriptions should feature close-up views of the subject's face.
> - Sensory details should enhance the scene (lighting, surroundings, sounds, smells, etc.), but keep the focus on the subject's face. The environment should be vivid but relevant to the subject's action or setting.
>
> 4. Tag the subjects:
>
> - Tag the subject with `<subj1> </subj1>`. For example,"a `<subj1>`man`</subj1>` stands in the room"
> - Do not tag pronouns (such as "he", "his", "him", etc.).
> - If there are multiple subjects, use different indexes for each individual (e.g., `<subj1>`, `<subj2>`, etc.).
> - Use the same index for all references to the same subject.
>
> 5. Length: No more than three sentences.
>
> Now generate 50 more samples with scenes related to technical or professional settings. For example, locations can be offices, schools, hospitals, labs, farms, factories, studios, kitchens, etc.
>
> Now generate 50 more samples with scenes related to casual, cultural, or recreational occasions, such as dances, music, dramas, movies, sports, arts, etc.
>
> Now generate 50 more samples with scenes related to outdoor activities or in nature, such as gardens, parks, mountains, forests, beaches, rivers, lakes, etc.

Prompts of Qualitative Comparison. The images in Figure 4 are generated with the following prompts:

- A man fixes his bicycle chain in the workshop, grease streaking his concentrated face. The smell of rubber and oil surrounds him as he works by lamplight.
- A man kneads bread dough in his sun-drenched kitchen, flour dusting his smile lines as morning light streams through gauzy curtains. His focused gaze follows the rhythmic movements of his hands, while the yeasty aroma fills the warm air.
- At the television studio, a man directs a live broadcast, speaking calmly into his headset. His focused gaze darts between multiple monitors as he calls camera changes.
- A man tracks wildlife in the misty forest, his experienced eyes reading subtle signs in the undergrowth. The early morning light filters through the canopy, illuminating his weathered face as he studies fresh prints.
- A woman tends to her rooftop beehives, moving with calm confidence among the buzzing insects. Her peaceful expression reflects years of experience as she checks each frame.
- Under fluorescent lights at the corner store, a woman browses magazine covers, her reflection ghosted in the glossy pages. Her fingers trace headlines as she squints slightly, the artificial brightness highlighting the fine lines around her eyes.
- In the acoustically treated recording studio, a woman masters audio tracks, her trained ear catching subtle nuances. Her eyes close briefly as she adjusts levels with precise movements.
- A woman fills her bird feeder in the backyard, morning dew soaking the hem of her robe. She squints against the rising sun, watching finches dart around her head as she pours the seeds.

Tackling Subject Ambiguity. Our experiments focus on single-subject consistency. To create a clean training dataset and minimize ambiguity, we use CelebA (Liu et al., 2015), which primarily contains human portraits. We generate masks using a powerful, off-the-shelf model, Grounding SAM 2 (Ren et al., 2024). We then manually filter out images with more than one face according to segmentation results. This step ensures that the training images contain single subjects and reduces ambiguous masks.

Extension to Multi-Subject Scenes. Although the model is trained only on single-subject portrait images, it can handle subject consistency in more challenging conditions, such as free-form prompts in *Scenes* and images with two subjects. For more challenging scenarios with multi-subject images, the proposed loss function framework can be further extended. Since the prior work (Hertz et al., 2022) has shown the relationships between word embeddings and feature maps, and a pseudo-word functions in the text embedding space to describe the subject that follows, we hypothesize that the model can learn to associate a pseudo-word with features of its corresponding subject, even in a multi-subject scene.

More thorough experimentation and evaluations need to be conducted for multi-subject scenarios, which will be part of our future work.

Segmentation Accuracy. In the single-subject experiments, the issue of inaccurate segmentation is less pronounced. The task of segmenting a person in a portrait, as is common in the CelebA dataset, is relatively straightforward for a powerful model like Grounded SAM 2. In addition, we have manually validated 100 randomly sampled images and find the predictions to be consistently reliable for this use case.

Licenses. The main experiments are conducted on CelebA (Liu et al., 2015). It is made available for non-commercial research purposes and requires users to comply with the terms outlined in the official usage agreement.

E Further Discussions

Preventing Prompt Corruption. To ensure that a pseudo-word only affects the subject and does not corrupt the rest of the prompt, we use a background preservation loss. This loss minimizes the difference in the background areas between an image generated with the pseudo-word and one generated without it. This localizes the effect of a pseudo-word to the masked subject area, thereby preserving the overall scene context dictated by the original prompt. The ablation study in Table 4 shows that removing this loss significantly harms prompt fidelity.

Place of Pseudo-Words. CoCoIns is developed based on the assumption that the pseudo-word lies in the text embedding space. Our mapping network f is explicitly designed to take a latent code $\boldsymbol{z}$ and output

a pseudo-word $\boldsymbol{w}$ as a vector that has the same dimension as the text embeddings of the pre-trained text encoder. This pseudo-word vector is then inserted directly into the sequence of prompt embeddings before the subject token.

Aligning Pseudo-Words. The consistency loss is not defined on the pseudo-words directly. Instead, a pseudo-word is made meaningful through an indirect alignment process using a contrastive loss on the generated images.

We use a triplet loss that operates on the predicted image latents from the denoiser. The model is trained to minimize the visual difference between images generated with the same pseudo-word (even with different prompts and noise) while maximizing the visual difference between images generated with different pseudo-words. This process forces the model to treat each pseudo-word as a unique identifier for a specific visual appearance. The meaning of a pseudo-word becomes the consistent subject identity it produces.

We also analyze the space of pseudo-words in Appendix C.3, which provides strong evidence that this indirect alignment is successful. The experimental results show that:

- **Interpolation**: Smoothly interpolating between two latent codes results in a smooth visual transition between the two corresponding faces.
- **Similarity**: Pseudo-words that are closer to each other in the embedding space (i.e., have higher cosine similarity) produce subjects with a more similar appearance.

These results demonstrate that the learned space of pseudo-words is well-structured and meaningful.

F Broader Impacts

Our approach enables image generative models to maintain subject consistency, extending their applicability across various tasks and modalities. Although this versatility could potentially be misused, such risks can be effectively mitigated through responsible deployment strategies, including strict usage policies, gated releases, and watermarking techniques.