

# HYPNOS : Highly Precise Foreground-focused Diffusion Finetuning for Inanimate Objects

Oliverio Theophilus Nathanael<sup>1</sup>, Jonathan Samuel Lumentut<sup>1</sup>, Nicholas Hans Muliawan<sup>1</sup>, Edbert Valencio Angky<sup>1</sup>, Felix Indra Kurniadi<sup>2</sup>, Alfi Yusrotis Zakiyyah<sup>1</sup>, and Jeklin Harefa<sup>1</sup>

<sup>1</sup> Universitas Bina Nusantara, Jakarta, Indonesia

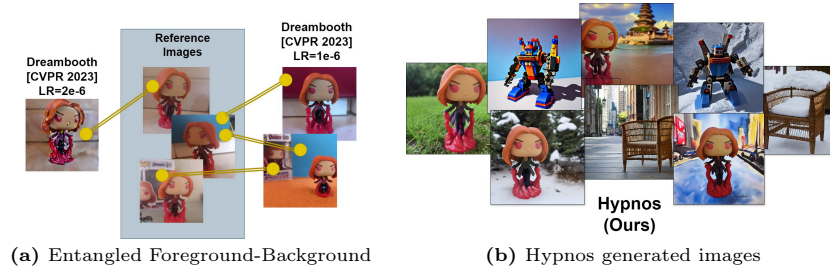
<sup>2</sup> Università di Pisa, Pisa, Italy

**Abstract.** In recent years, personalized diffusion-based text-to-image generative tasks have been a hot topic in computer vision studies. A robust diffusion model is determined by its ability to perform near-perfect reconstruction of certain product outcomes given few related input samples. Unfortunately, the current prominent diffusion-based finetuning technique falls short in maintaining the foreground object consistency while being constrained to produce diverse backgrounds in the image outcome. In the worst scenario, the overfitting issue may occur, meaning that the foreground object is less controllable due to the condition above, for example, the input prompt information is transferred ambiguously to both foreground and background regions, instead of the supposed background region only. To tackle the issues above, we proposed Hypnos, a highly precise foreground-focused diffusion finetuning technique. On the image level, this strategy works best for inanimate object generation tasks, and to do so, Hypnos implements two main approaches, namely: (i) a content-centric prompting strategy and (ii) the utilization of our additional foreground-focused discriminative module. The utilized module is connected with the diffusion model and finetuned with our proposed set of supervision mechanism. Combining the strategies above yielded to the foreground-background disentanglement capability of the diffusion model. Our experimental results showed that the proposed strategy gave a more robust performance and visually pleasing results compared to the former technique. For better elaborations, we also provided extensive studies to assess the fruitful outcomes above, which reveal how personalization behaves in regard to several training conditions.

**Keywords:** Generative model · Stable Diffusion · Dreambooth

## 1 Introduction

Personalization for text-to-image (T2I) Diffusion Model has been a hot field of study and rapidly popularized among researchers in computer vision studies as well as practitioners and even hobbyist. Personalization is the key for generative



**Fig. 1:** The current (a) widely used method [21] is prone to subject-scene entanglement and overfitting as shown by the yellow lines. Our proposed method visually (b) shows a realistic disentanglement effect between subject and scene at the foreground and background, respectively.

models to be widely used as a tool for various use cases. This personalization in the Diffusion Model has the same spirit as grounding Large Language Models (LLMs), where it opens the possibility of being reliably incorporated in diverse real-world environments. With reliable personalization in the Diffusion Model, it opens several potentials for recreational purposes and even in the business world, such as creating product advertisements and campaign photos, which mostly highlight the foreground objects (*foreground-focused objective*).

One particular distinctive family of these techniques, namely Dreambooth [21], enables the ability to finetune Diffusion models on a specific image without losing prior knowledge. At the moment Dreambooth is one of the most popular personalization techniques because of its efficiency and minimum requirements of input samples (3-5 reference images) to be able to synthesize great quality, with high similarities to the reference input.

Despite that, there is still room for improvement, as the current method is still prone to structure and color distortions, which are mostly caused by the coarse attempt of the model to blend the object in the foreground and the scene in the background [21]. Visually, there is an entanglement of background and foreground semantic information as shown in Fig. 1a (content relations are highlighted with yellow lines). To overcome this issue, one might increase the learning rate of [21] to enforce more consistency to the reference images. Unfortunately, such approach causes another problem to arise [21], which is the overfitting issue.

To our observation, the work of [21] tends to produce the entangled version of both foreground and background outputs from the given foreground reference input (Fig. 1a). Adjusting the learning rate to the higher (or lower) values in [21] finetuning affects the variance of both foreground and background information. This can be perceived as an entangled foreground-background issue in the T2I task, which is improper for the foreground-focused objective mentioned above.

In our pursuit of creating a reliable T2I finetuning technique, we introduce HYPNOS, a highly precise foreground-focused diffusion finetuning approach for inanimate objects. We propose several strategies that work synergically to disentangle foreground-background information. These additions are proven to pro-

duce visually consistent and pleasing images as seen on Fig. 1b. In detail, our strategies comprises of: straightforward image dataset augmentation, explicit foreground-background prompts procedures, and sets of new supervision mechanisms to quantify the foreground deviation from the reference images. Importantly, those method are implemented while still maintaining fast finetuning time and low latency architecture. We also conducted an extensive evaluation of both qualitative and quantitative analysis to further support our studies. These evaluations are also used to explain some distinct behaviors of Hypnos compared to Dreambooth and Textual Inversion techniques. Our main contributions can be summarized as follows:

- We propose a novel technique with competitive result by leveraging content-centric augmentation and new sets of supervision mechanisms. This improvement reliably ensures foreground-background disentanglement, significantly lowers noise, and enables semantic level tuning for the T2I task.
- We introduce prompt-invariant and prompt-varying metrics as a new approach to quantitatively assess the Dreambooth family technique and how to interpret the results. We show that these new metrics can be used to gain a better insight of the finetuned model.
- We conduct analysis and experiments on the introduced hyperparameters to elaborate further on their effects and how to tune them properly.

## 2 Related Works

### 2.1 Vision-based Generative Model

Diffusion Model is a groundbreaking image generation method that leverages thermodynamics based process [27]. It was then further popularized by the introduction of the Denoising Diffusion Probabilistic Model [8]. It is a probabilistic model that is capable of modeling a complex distribution using a denoising process. Denoising Diffusion Implicit Model later eliminated the markovian assumption which speeds up the denoising process [28]. Other popular improvement includes the adaptation of Cosine scheduling on the noise diffusion process to ensure the image is not too rapidly destroyed throughout the diffusion process [15]. Other physics-inspired generative models include Poisson Flow Generative Models (PFGM), which is inspired by the physics of Electrodynamics [30]. Its improvement, PFGM++ [31], unified PFGM Model with Diffusion Model.

The current trend involves the strategy of text-prompt-based input to guide the denoising process, hence called as T2I diffusion model. CLIP, a transformer-based model that aligns text and image embedding [18], is introduced to satisfy such task. By utilizing CLIP, The work of GLIDE [14] showed to produce great quality results. Recent text transformer-based works, e.g. BERT [4] and T5 [19], also showed great performance in running the image generative task [23].

Diffusion process recently involved the denoising procedure on Variational Autoencoder’s [10] latent space to make the model resource-effective. This method is commonly known as the Latent Diffusion Model [20]. The following are among

the popular latent-based works: Stable Diffusion [20], Stable Diffusion XL [16], Dall-E 3 [26], and the recent video-based (temporal) diffusion generative task [1].

## 2.2 Personalization

As explained on Sec. 1 personalization aims to enable T2I models to be able to generate a specific object based on some reference images. Some of the methods include textual inversion [6] and controlnets [32]. Among many personalization techniques, the one that arguably stands out the most is Dreambooth [21]. Dreambooth finetunes a diffusion model using 3-5 reference images and some prior class images to prevent the loss of prior ability to produce diverse images on that particular class. Dreambooth also can be trained relatively fast.

In recent years, numerous works have been proposed to better fit Dreambooth in some use cases. LoRA [9] speeds up the finetuning process at the expense of image quality. Google’s HyperDreamBooth [22] focuses on fast personalization for human faces using Hypernetworks to adjust main network weights [3]; further accelerated by LoRA. DreamCom [12] had shown remarkable results on the image inpainting task by reusing some good generated images to retrain the model. Other variations of Dreambooth include the procedure of learning other concepts besides the object appearance itself [13]. In this work, we opt to scope the usage of Dreambooth [21] (comprises of Stable Diffusion [20]) as the baseline for our foreground-focused diffusion finetuning task on inanimate objects.

## 3 Methodology

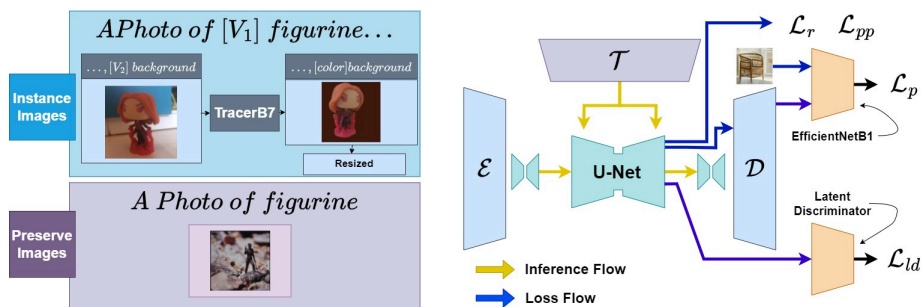
The proposed method is crafted to accommodate Dreambooth in terms of producing a more consistent foreground subject. Hence, we utilized the same backbone model, but with additional proposed functionalities to better disentangle the foreground-background information. As seen on Fig. 2, there are two main part of the process. The *first one* is the datasets refinement procedure that undergoes our specific image augmentation and prompt engineering functionalities (left side of Fig. 2). The *second part* is the Diffusion model itself that is trained on our diverse supervision mechanisms (right side of Fig. 2).

### 3.1 Dataset Refinement

Similar to Dreambooth, Hypnos is a finetuning technique that only need 3-5 image samples. We limit the scope of the generation task to only for inanimate objects. Our images that are used within this work are taken from personal images and Unsplash (<https://unsplash.com/>).

### 3.2 Text to Image Model

To generate the image we use a stable diffusion v1.5 model [20]. The model itself comprises of several networks that work synergically. These models usually utilize U-Net as the network to model the denoising process.



**Fig. 2:** We proposed to apply image augmentation and prompt engineering to the instance dataset, we also apply sets of losses as shown by the blue arrows

Stable Diffusion is a latent diffusion model which means that the diffusion and denoising process happens on a latent representation of an image. Furthermore, as a Text to Image model, it also incorporates CLIP embeddings into the U-Net model to guide the generated image based on a specific prompt. Following the common procedure, the images in our method are encoded by the Variational Autoencoders (VAE), while the text prompts are processed by the CLIP method.

In terms of finetuning, there are several approaches to run such task using the Dreambooth method. In this work, we opt to freeze the weight of the VAE, meaning that Hypnos only optimize both the denoising model and the text encoder. We also preserve the denoising procedure from the Stable Diffusion which maintain the lower latency characteristic [20].

### 3.3 Image Augmentation and Prompt Engineering Functionalities

Based on the original Dreambooth paper there are two sets of dataset to be extracted for supervision (Fig. 2 (left)). The *first one* (purple region of Fig. 2 (left)) is the self generated class images paired with standard prompt such as "a photo of [C]" where [C] is the class name of the subject. The *second set* (blue region of Fig. 2 (left)) is the specific provided subject images and traditionally labeled as "a photo of [V] [C]" where [V] is a rare word that acts as a name to tell our object apart from others. Both extracted datasets are utilized in supervision to prevent the lost of the prior general understanding of that particular class.

Unfortunately, doing the above’s strategy in Dreambooth [21] invokes the entanglement effect in both foreground and background, which is unfavorable for foreground-focused objective. To avoid this, we propose a content-centric augmentation that explicitly introduces background-foreground disentanglement. We then utilized TracerB7 [11, 17] to randomly replace the background with monotone colors. We empirically set the augmentation proportion to 0.66 to avoid losing the ability to blend the subject with the background. To further introduce the foreground diversity, we also resized a small portion of the sampled dataset spatially. Moreover, we provided *prompting adjustment strategy* to

allow the model to recognize the foreground and background via first and second clauses, respectively, as seen on Fig. 2 To describe the original image’s background, it is impractical to use a particular color name as used in the background changed images; hence, we opted to utilize another rare word as a placeholder.

### 3.4 Supervision Mechanisms

One of our proposed method includes specialized losses to make sure that the model will preserve the foreground structure and color while still letting the background to exhibit a high variance. Our proposed method is a combination of 4 different losses which are reconstruction losses, prior preservation loss, perceptual loss, and latent discriminator loss as shown with the blue arrows on Fig. 2, which is written as Eq. (1):

$$\mathcal{L} = \lambda_r \mathcal{L}_r + \lambda_{pp} \mathcal{L}_{pp} + \lambda_p \mathcal{L}_p + \lambda_{ld} \mathcal{L}_{ld}. \quad (1)$$

Based on experiments it is generally best to set  $\lambda_r = 1$ ,  $\lambda_{pp} = 1$ ,  $\lambda_p = 0.003$ ,  $\lambda_{ld} = 0.5$ . Note that these weights itself are hyperparameters, hence it can be adjusted based on the intended model output.

**Reconstruction Loss ( $\mathcal{L}_r$ )** Traditionally this loss is simply the mean squared error of the predicted noise of the instance latent images which denotes by  $z_i$  at timestep  $t$  where  $t \sim \mathcal{U}(1, 1000)$ . This loss ensures the generated image is as similar as possible to the instance image. Though mean squared error is a widely known loss for reliably quantifying deviation from a target value, our observation suggests that exponential-growth-base loss function is preferable than quadratic-growth-base one. This is done as measuring the deviation from a noise on a latent space is focused on the data with normal distribution. One intuitive alternative is to use the inverse of gaussian distribution as shown on Eq. (2) as it is based of the VAE overall distribution itself.

$$\mathcal{L}_r = \sigma \sqrt{2\pi} (\mathbb{E}_{z_i, t, c, \epsilon} \|e^{\|\epsilon_\theta(z_i, t, c) - \epsilon\|_2^2 / 2\sigma^2} - 1) \quad (2)$$

Based on our experiment, this is very effective for getting rid of the noise artefact that is often seen in a higher learning rate of Dreambooth method (details on *Supplementary Material*). By adjusting the  $\sigma$  it is possible to tune on how steep the loss is. In detail, the L2 loss can be re-approximated by tweaking  $\sigma$ , particularly by minimizing the least squared regression of both function for bound  $\pm a$  as shown on Eq. (3).

$$\min_{\sigma} \int_0^a (x^2 - \sigma \sqrt{2\pi} (e^{x^2/2\sigma^2} - 1))^2 dx \quad (3)$$

By solving Eq. (3) for  $a$  equals to 1, where reconstruction loss are very unlikely to exceed 1, it is empirically found that  $\sigma = 1.382$  is the closest to L2 loss, with standard deviation larger than that, meaning that the loss is flatter than L2 loss.

**Prior Preservation Loss ( $\mathcal{L}_{pp}$ )** While Reconstruction Loss computes the deviation between the generated image with respect to the subject latent images, Prior Perceptual loss works with respect to the class latent images which denotes by  $z_p$  as written in Eq. (4):

$$\mathcal{L}_{pp} = \mathbb{E}_{z_p, t, c, \epsilon} \|\epsilon_{\theta}(z_p, t, c) - \epsilon\|_2^2. \quad (4)$$

In this work, we opt to preserve the original MSE loss on this prior preservation function, as this allows the model to focus its learning process on the instance images rather than the class images.

**Perceptual Loss ( $\mathcal{L}_p$ )** Perceptual Loss is widely used for style transfer task. The advantage of using this loss is that it can semantically quantify the deviation of the generated image. We utilized EfficientNetB1 [29] (a classifier network) as the encoder and take the L2 loss across several of its activation. To preserve low-level semantic information, we gave larger weight to the activations of its shallow layers Furthermore, using classifier network to guide a diffusion model is a common practice. One justification includes the use of classifier output to aid an ambiguous text to produce an image that is better represent the intended picture [25].

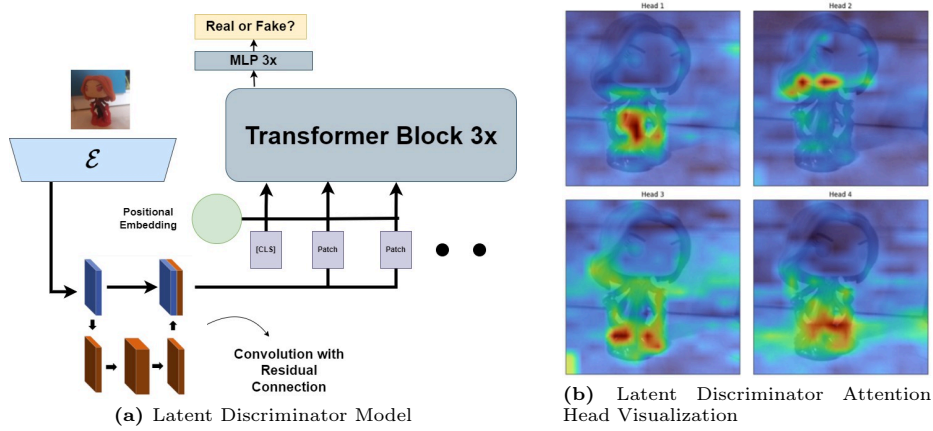
Since classifier networks process decoded denoised images, thus, to be able to apply perceptual loss, it is required to denoise the images straight to  $z_0$  and then decode the image beforehand. The resulting equation is described on Eq. (5):

$$\mathcal{L}_p = \mathbb{E}_{z_i, t, c, x} \|C_{\theta}(x) - C_{\theta}(D_{\theta}(z_0))\|_2^2, \quad (5)$$

where  $C_{\theta}$  is a classifier model,  $D$  denotes the VAE Decoder,  $x$  denotes the reference image, and  $z_0$  as the denoised latent.

Based on our observation, to avoid overfitting, there are two approaches that can be done, *the first one* is to apply Latent Discriminator that balances the loss and *second one* is by limiting its influence on the loss, rather than decreasing the loss weight. The latter can be implemented by limiting the number of steps that are influenced by Perceptual Loss. Intuitively, this is done to avoid local minima of the original Dreambooth loss that exhibit a high perceptual dissimilarity as described on Fig. 1a. It is also important to avoid a perfect perceptual match. Therefore, Perceptual Loss should guide optimization for only a certain number of steps, with the remaining steps optimized without it. To do so, we introduced a new hyperparameter  $s_p$  that directly control the perceptual similarity of the generated image by limiting the maximum step influenced by Perceptual loss. We set  $s_p = 500$  out of 800 steps of total training cycle on the re-defined Perceptual loss  $\mathcal{L}_p$  in Eq. (6).

$$\mathcal{L}_p = \begin{cases} \mathbb{E}_{z_i, t, c, x} \|C_{\theta}(x) - C_{\theta}(D_{\theta}(z_0))\|_2^2, & \text{if } s \leq s_p \\ 0, & \text{otherwise} \end{cases} \quad (6)$$



**Fig. 3:** Latent Discriminator is designed to be lightweight and has low latency. Despite the small size, the discriminator successfully identify important feature as shown by the attention head visualization

**Latent Discriminator Loss ( $\mathcal{L}_{ld}$ )** Similar to perceptual loss, this loss aims to quantify the foreground deviation semantically to preserve the foreground information. Rather than computing the loss of the decoded image, this loss computes the difference straight from the denoised latent space. To achieve this, we make a discriminator trained on the latent representation of the instance image and the monotone colored background version as the real images while fake images comprises of class images and altered instance images such as removed foreground and negative colored foreground (details on *supplementary material*). The discriminator itself is a 3 layer vision transformer [5] with additional convolutional layers to ensure efficient semantic information extraction (visualized in Fig. 3a).

Similar to Adversarial Diffusion Distillation [24] that applied adversarial learning on diffusion model, our latent discriminator is also provided with the flexibility to compete with main model. It is written in Eq. (7) with  $LD_{\theta}(\cdot)$  and  $z_0$  denote the Latent Discriminator model and denoised latent image, respectively:

$$\mathcal{L}_{ld} = \mathbb{E}_{z_i, t} \|1 - LD_{\theta}(z_0)\|_2^2. \quad (7)$$

The initial intention of this loss ( $\mathcal{L}_{ld}$ ) is to provide the diffusion model with an explicit semantic similarity loss referenced to all instance images. By this we can utilize all image references rather than just one sample as in the other Losses ( $\mathcal{L}_r$ ,  $\mathcal{L}_{pp}$ ,  $\mathcal{L}_p$ ). Thus, adversarial learning can be viewed as an effort for the discriminator model only to adapt with the everchanging image quality of the generated image. We design our discriminator with light architecture and train it initially with 600 steps, and then followed by the combined training.

The implemented architecture for Latent Discriminator is shown on Fig. 3a. As seen on Fig. 3b, this simple architecture is enough to detect important features of the object.



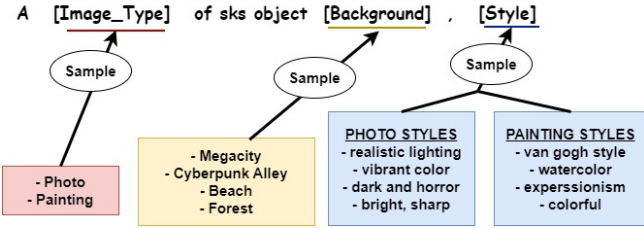


Fig. 4: Prompts for Varying Prompt Evaluation are sampled from the combination of predefined list for image type, background, and style

### 3.5 Evaluation Metrics

Quantifying Dreambooth method family is one of the most tricky task to handle there are at least two main arguments to support this claim. The first one is that Dreambooth family falls into few-shot learning category which by definition has a scarce amount of ground truth, hence it is not suitable to represent the approximate intended data distribution, rather it is expected to be able extrapolate from that given data distribution. This imply that the measure of how good the extrapolation result vary a lot across preferences and use cases. Second, ground truth incorporating diverse prompt are non-existent, unlike what seen in classic Diffusion model task or even controlnet model [32]. This leads to some evaluation blind spot on how well model can adapt to diverse prompts and how does the model adapt to the prompt and what are the impact to the object.

**Prompt Invariant and Prompt Varying Evaluation** In this work, we proposed a modification to the existing metric evaluation to open new perspective and used it alongside the existing method. This is reasonable since Dreambooth family techniques usually use standard prompt to generate all the images which we called as *Prompt Invariant Evaluation*. To better accomodate the current evaluation approach, we propose that it is also important to assess the adaptibility of the resulting model on diverse prompt, which we coined as *Prompt Varying Evaluation*.

To minimize bias that emerges while handcrafting every single prompt, we proposed a solution that samples parts of a generic image generation prompt structure as seen on Fig. 4. There are three main considerations, the first one is the *image type* such as photo or painting, the second part is the *background information*. In this context, background information is not only limited on the background scene but also the image composition other than the main object itself. The third one is the *styles*; style descriptions are crucial to generate high quality image in smaller diffusion models as they require longer prompt compared to larger model such as SDXL [16]. Based on our understanding, style descriptions may vary across image types, hence its description is acquired from the previously sampled image type.

**Table 1: *Prompt Invariant*** quantitative metrics evaluated on 3 datasets, Funko figurine (●), Rattan chair (●), and Lego Robot (●).

Method	DINO	CLIP-I	CLIP-T	FID	SSIM	PSNR	LPIPS
Hynos (Ours)	● <b>0.7851</b>	<b>0.8635</b>	0.0094	3,6032	0.5974	11.8504	0.3850
	● <b>0.6502</b>	<b>0.8015</b>	0.0067	<b>2.3840</b>	<b>0.2225</b>	<b>9.4634</b>	<b>0.4166</b>
	● 0.6589	0.8369	0.0183	<b>5.6330</b>	0.3876	<b>10.5387</b>	0.4624
Dreambooth (LR=1e-6)	● 0.6422	0.7935	0.0183	<b>2.9873</b>	<b>0.6056</b>	<b>12.2883</b>	<b>0.3663</b>
	● 0.5012	0.7404	<b>0.0549</b>	13.1933	0.1645	9.0604	0.4583
	● <b>0.7130</b>	<b>0.8753</b>	<b>0.0458</b>	5.7367	0.3429	9.3005	0.4679
Dreambooth (LR=2e-6)	● 0.5311	0.7468	0.0153	14.7671	0.4781	11.4756	0.4513
	● 0.2647	0.4742	0.0224	42.7634	0.1433	9.3789	0.5128
	● 0.5704	0.8323	0.0175	14.2261	0.3060	9.3813	<b>0.4622</b>
Textual Inversion	● 0.4934	0.6469	<b>0.0417</b>	12.2159	0.4565	9.9478	0.4917
	● 0.4397	0.7134	0.0308	4.6512	0.2125	8.8142	0.4785
	● 0.3904	0.6118	0.0312	6.4942	<b>0.3929</b>	9.6875	0.5160

**Dreambooth Metrics** As perceived from the Dreambooth work, their finetuning technique is a unique task that is inadequate to be measured using classic image generation metrics because of its nature that have limited reference image as explained earlier. Following their scheme, we use DINO [2], CLIP-I, and CLIP-T to perform the baseline quantitative measurement.

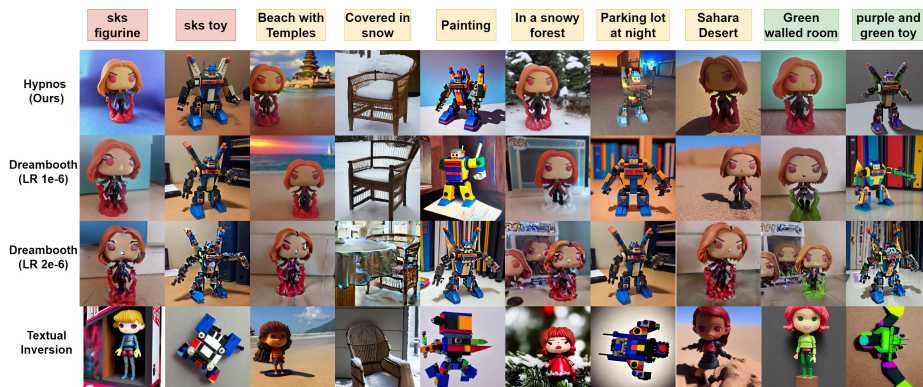
**Image Quality and Perceptual Metrics** Apart from the metrics above, image generation tasks are also evaluated using metrics such as FID [7], LPIPS [33], SSIM, and PSNR. These metrics are rarely used to evaluate the Dreambooth method family as it is prone to biased score. However, we still opt to include them as they still can provide additional insight on the generated image quality.

## 4 Result and Discussion

### 4.1 Evaluation

As mentioned on Sec. 3.5, the evaluation of Dreambooth method family can vary across preference and use cases. Therefore, we conduct both qualitative and quantitative analysis of the results. As shown later on this section, solely analyzing quantitative metrics on this particular task has a potential to be misleading, hence it is advised to interpret the quantitative assessment with a grain of salt. We resolve the issues above by employing the procedure of always combining quantitative and qualitative analysis throughout the evaluation process.

**Prompt Invariant Evaluation** As seen on Tab. 1, we evaluated each metrics on three subject samples which are a Funko figurine (●), a rattan chair (●), and a Lego robot (●). This approach can be observed on Tab. 1. We opt to experiment on these samples to minimize the bias effect while assessing the metrics. The prior method of Dreambooth utilized a small number of instance



**Fig. 5:** Image generation comparison, red prompt denotes *prompt invariant*, yellow prompt denotes *prompt varying*, green prompt denotes specific prompting to analyze foreground-background disentanglement ability and highlight semantic leaking

images that were insufficient to be referred to as the perfect target distribution. This yields the metrics results that vary vastly across datasets, which we desire to avoid. By this reasoning, we performed a comparison if and only if those methods were trained with the same instance images.

As shown in Tab. 1, Hypnos is consistently superior in DINO and CLIP-I metrics. In contrast, Dreambooth mainly dominates CLIP-T. However, CLIP-T’s metric is the least aligned with qualitative outcomes (refer to CLIP-T scores in Tab. 1 vs. results in Fig. 5). We suspect that this is caused by the large domain gap between post-training CLIP text embedding and pre-trained CLIP image embedding, which, most of the time, each embedding deviates and becomes unreliable. DINO and CLIP-I, on the other hand, are rather in line with what was observed qualitatively. Visually, Hypnos exhibits minimum noise effect in the generated images while preserving most foreground information similar to the dataset. Dreambooth, even with a lower learning rate case, produces high-noise results (*first and second column* in Fig. 5) and tends to overfit on the instance image. This is reasonable because both the image and prompt are what exactly the model learned in the training phase. In contrast, Hypnos introduced higher variance in the training background with its respective prompt, resulting in arbitrary background results when no background information is provided in the prompt input. At the same time, Hypnos is still able to maintain structurally similar foreground outputs (*first and second column* in Fig. 5).

Based on our observation, Dreambooth tended to yield high noise when the object or scene comprises of mostly homogeneous regions (e.g. figurine case). On the Lego robot case, Dreambooth gave perfect generation results although lower learning rate is utilized and thus it produced higher scores in the metrics. Nonetheless, the noise level becomes aggravated in the Lego case if a higher learning rate is employed. Hypnos tackles the above problem as shown in the first row of Fig. 5. In the case of longer prompts input (yellowish prompt in Fig. 5),

**Table 2: *Prompt Varying*** quantitative metrics evaluated on 3 datasets, Funko figurine (●), Rattan chair (●), and Lego Robot (●).

Method	DINO	CLIP-I	CLIP-T	FID	SSIM	PSNR	LPIPS
Hypnos (Ours)	● <b>0.7070</b>	0.7883	0.0200	11.0675	0.5139	10.9563	0.4626
	● <b>0.5461</b>	<b>0.6572</b>	<b>0.0326</b>	<b>8.6402</b>	<b>0.1797</b>	8.8435	0.5039
	● 0.4920	0.6462	0.0242	22.5143	0.2863	<b>9.7863</b>	0.5392
Dreambooth (LR=1e-6)	● 0.7050	0.7837	0.0224	6.5111	0.5453	10.7109	0.4402
	● 0.4499	0.5814	0.0173	11.2734	0.1687	8.1098	0.5196
	● 0.4377	0.6685	0.0286	14.6446	0.2887	9.1872	0.5502
Dreambooth (LR=2e-6)	● 0.6630	<b>0.8028</b>	0.0204	<b>5.1134</b>	<b>0.5589</b>	<b>11.6786</b>	<b>0.4089</b>
	● 0.4656	0.6424	0.0179	17.1525	0.1650	<b>9.1083</b>	<b>0.4583</b>
	● <b>0.5826</b>	<b>0.7704</b>	<b>0.0336</b>	<b>10.4226</b>	<b>0.3325</b>	9.7451	<b>0.4830</b>
Textual Inversion	● 0.3131	0.4355	<b>0.0297</b>	42.3451	0.3066	8.9635	0.5942
	● 0.3242	0.5427	0.0224	18.9589	0.1273	7.8409	0.5455
	● 0.3132	0.5051	0.0239	25.5648	0.2397	8.6579	0.5935

unlike Dreambooth, Hypnos avoids the overfit and underfit effects, which yields the production of highly similar and high-quality foreground results.

Furthermore, as seen on Fig. 5 (*fourth row*), Textual Inversion produced a poor output in terms of similarity even with the standard prompt usage. From this point onward, Textual Inversion acted as a baseline and control to our metrics analysis as it is already widely known to underperform compared to the classic Dreambooth.

Tab. 1 also shows standard image generation metrics with the prompt invariant case. The metrics provided useful insights, yet they are insufficient to determine model superiority. Nevertheless, Hypnos still dominates most of the metrics. In terms of FID, Hypnos consistently generated a perceptually similar image to the instance images. This is further supported by higher SSIM and PSNR scores (highlighted with **bolds** in Tab. 1; **bolds** indicate best scores).

**Prompt Varying Evaluation** Tab. 2 shows the metrics score across varying prompts. Varying prompts is expected to force the models to produce diverse images. This implies that the similarity index should decrease significantly. Those performances are observed in Hypnos when comparing the scores in Tab. 2 and Tab. 1. A noticeable change is that the higher learning rate Dreambooth dominates most metrics, suggesting high similarity. This phenomenon can be explained by observing the resulting images as seen on the *third row* in Fig. 5. Qualitatively, it is apparent that Dreambooth (LR=2e-6) is overfitted, and the generated images fail to adapt to the prompt. Thus, Dreambooth’s high scores are obtained from overfitting, which is inadequate.

On the other hand, Dreambooth, with a lower learning rate, seems to be able to better adapt to the prompt. However, as mentioned earlier, they came with a downside effect on the foreground quality. As seen on the *fifth* and *eighth column* of Fig. 5, it is capable of adapting to the prompt input, but its foreground object is highly altered. Hypnos, in contrast, exhibited a more consistent and robust foreground similarity.

Another downside of Dreambooth is that it is often overfit on the background. As seen on the *third column* of Dreambooth case of Fig. 5, the beach looks like a room, and on the *sixth column*, Dreambooth fails to produce a forest scene. It instead overfits the dataset background (*second row* of Fig. 5). Another example is shown on the *fifth* and *seventh column* where the toy clearly always stands in front of books and fails to follow the prompt; this is because some of the training images are taken in front of the books. These issues are effectively resolved on Hypnos-generated images as they are trained to have content-centric knowledge.

Hypnos also displayed a clear distinction between foreground and background. This statement is justified by how it preserves the image color and adapts to the scene naturally. On the other hand, Dreambooth showed the foreground-background entanglement effect, in which its phenomenon is observed on the *sixth* and *eighth column* of Fig. 5, where the base of the figurine’s colors are affected by the scenes’ color.

Furthermore, Hypnos’ robust foreground structure preservation allows it to produce more imaginative images, such as snow on a rattan chair (*first row fourth column* of Fig. 5). In that case, both components are rarely seen together in the real world. Dreambooth, however, generated a very different chair and struggled with high-frequency features like a rattan chair (*fourth column* of Fig. 5). This explains why Hypnos dominates the metrics for Rattan Chair as seen on both tables (Tab. 1 and Tab. 2).

**Foreground-Background Disentanglement Analysis** Hypnos also solved another limitation of Dreambooth [21], mentioned in their work, which is the color and theme of the scene that leaks information from the foreground object or vice versa (entangled). As seen on Fig. 5 with green prompts, Dreambooth produced a highly overfitted image with noises when prompted with green walled room. Rather than coloring the room wall green, it painted the base green instead, and this phenomenon is frequently observed on Dreambooth. This phenomenon is also clearly observed in textual inversion, where the clothes are green instead of the wall. By leveraging content-centric strategies, Hypnos, in contrast, can correctly make a green walled room background while still having accurate foreground color (refer to the second column from right in Fig. 5).

Besides excelling in background modification, Hypnos also excels in foreground modification. This is observed to be challenging for Dreambooth, especially for the more complex objects where they often end up overfitting even with a lower learning rate (first column from the right of Fig. 5). Dreambooth structural consistency dropped with longer prompt and having difficulty adapting to the prompt whereas Hypnos successfully altered the object color to purple and green without altering much of the structure. This again proves the robustness of Hypnos in generating inanimate objects in various scenes and prompts.

## 4.2 Ablation Studies

**Perceptual Loss** We evaluate Hypnos on two configurations to analyze the effect of Perceptual loss. The First one is when we exclude any limit to the max-

imum step influenced by Perceptual Loss. In this case, it tends to highly overfit, such as the same lighting and pose information directly pasted from the original image. We also observed the emergence of background overfitting because of this uninformed subject-scene loss. Secondly, we excluded the Perceptual Loss. Without perceptual loss, structural preservation is downgraded and some noise started to showed up although still far less than Dreambooth since Hypnos also utilized inverse gaussian reconstruction loss. Refer to suppl. material for details.

**Latent Discriminator Loss** By removing the Latent Discriminator, the color consistency and background generation are severely impacted. This is reasonable as two-third of the training image scenes are just monotone colors, hence making the generated image have bad subject-scene blending.

### 4.3 Limitations

As shown on Fig. 5, Hypnos does not guarantee perfect image reconstruction nor 3D consistency. Hypnos also never intended to be used on humans nor animals hence it may not be optimized for those tasks. There also seems to be a tradeoff between high foreground reconstruction and subject-scene blending capability.

### 4.4 Societal Impact

Our method enables small businesses, up to the well-established businesses, to easily gain access to a more reliable and flexible high-quality product image generation for marketing purposes compared to background image inpainting as the most common approach at the time of writing of this work. On the other end of the spectrum, Hypnos can also potentially produce misleading images without any reality backing, which may lead to unintentional or even deliberate fraud or other malicious activities.

## 5 Conclusion

We propose Hypnos, a content-centric personalization diffusion finetuning technique that addresses high noise and foreground-background entanglement in T2I tasks, which are prevalent problems of its predecessors. Given 3-5 image input samples, Hypnos consistently preserves the foreground’s structure and color while enabling diverse background modification. We also show that it is also possible to adjust the diversity of the generated image semantically, with inanimate object constrain, by utilizing our proposed supervision mechanisms. Furthermore, with the newly introduced hyperparameters, Hypnos offers a wider space of adjustment on the target distribution to support diverse use cases. We believe that the development of Hypnos finetuning strategy paves the way for various foreground-focused downstream computer vision-based generative tasks.

**Acknowledgment** This research project is supported by the grant of Penelitian Pemula Binus No: 069A/VRRTT/III/2024 (Universitas Bina Nusantara).

## References

1. Blattmann, A., Dockhorn, T., Kulal, S., Mendeleevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023) [4](#)
2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021) [10](#)
3. Chauhan, V.K., Zhou, J., Lu, P., Molaei, S., Clifton, D.A.: A brief review of hypernetworks in deep learning. arXiv preprint arXiv:2306.06955 (2023) [4](#)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) [3](#)
5. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale (2021) [8](#)
6. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022) [4](#), [17](#)
7. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017) [10](#)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020) [3](#)
9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021) [4](#)
10. Kingma, D.P., Welling, M., et al.: An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning* **12**(4), 307–392 (2019) [3](#)
11. Lee, M.S., Shin, W., Han, S.W.: Tracer: Extreme attention guided salient object tracing network (student abstract). In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 12993–12994 (2022) [5](#)
12. Lu, L., Zhang, B., Niu, L.: Dreamcom: Finetuning text-guided inpainting model for image composition. arXiv preprint arXiv:2309.15508 (2023) [4](#)
13. Motamed, S., Paudel, D.P., Van Gool, L.: Lego: Learning to disentangle and invert concepts beyond object appearance in text-to-image diffusion models. arXiv preprint arXiv:2311.13833 (2023) [4](#)
14. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021) [3](#)
15. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International conference on machine learning. pp. 8162–8171. PMLR (2021) [3](#)
16. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023) [4](#), [9](#)
17. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jagersand, M.: U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition* **106**, 107404 (2020) [5](#)

18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) [3](#)
19. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* **21**(140), 1–67 (2020) [3](#)
20. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022) [3](#), [4](#), [5](#)
21. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023) [2](#), [4](#), [5](#), [13](#), [17](#)
22. Ruiz, N., Li, Y., Jampani, V., Wei, W., Hou, T., Pritch, Y., Wadhwa, N., Rubinstein, M., Aberman, K.: Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6527–6536 (2024) [4](#)
23. 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.: Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* **35**, 36479–36494 (2022) [3](#)
24. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042* (2023) [8](#)
25. Schwartz, I., Snæbjarnarson, V., Chefer, H., Belongie, S., Wolf, L., Benaim, S.: Discriminative class tokens for text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22725–22735 (2023) [7](#)
26. Shi, Z., Zhou, X., Qiu, X., Zhu, X.: Improving image captioning with better use of captions. *arXiv preprint arXiv:2006.11807* (2020) [4](#)
27. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015) [3](#)
28. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020) [3](#)
29. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019) [7](#)
30. Xu, Y., Liu, Z., Tegmark, M., Jaakkola, T.: Poisson flow generative models. *Advances in Neural Information Processing Systems* **35**, 16782–16795 (2022) [3](#)
31. Xu, Y., Liu, Z., Tian, Y., Tong, S., Tegmark, M., Jaakkola, T.: Pfgm++: Unlocking the potential of physics-inspired generative models. In: International Conference on Machine Learning. pp. 38566–38591. PMLR (2023) [3](#)
32. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023) [4](#), [9](#)
33. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) [10](#)