# HYPNOS : Highly Precise foreground-focused diffusion finetuning for inanimate objects

Oliverio Theophilus Nathanael[1] , Jonathan Samuel Lumentut[1] , Nicholas Hans Muliawan[1] , Edbert Valencio Angky[1] , Felix Indra Kurniadi[3] , Alfi Yusrotis Zakiyyah[2] , Jeklin Harefa[1]

[1]Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia
[2]Mathematics Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia
[3]Università di Pisa, Pisa 56126, Italy

This research project is supported by the grant of Penelitian Pemula Binus No: 069A/VRRTT/2023 (Universitas Bina Nusantara).

ACCV HANOI VIETNAM 2024 DEC 8-12

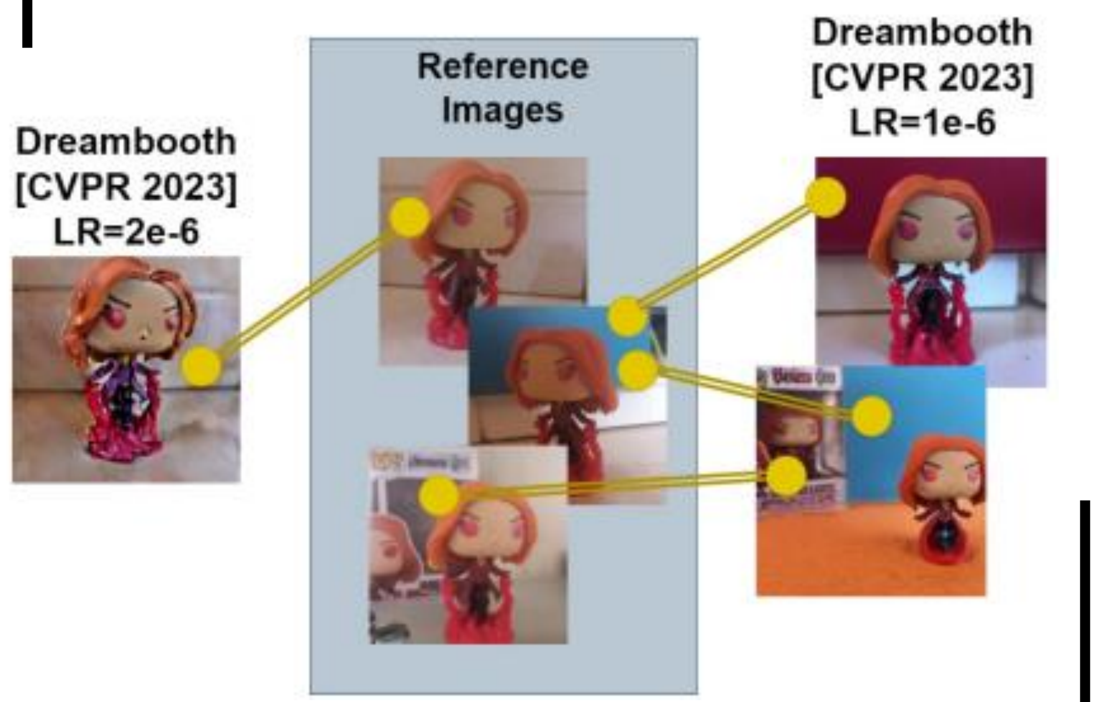
BINUS UNIVERSITY

UNIVERSITÀ DI PISA

## Introduction

### Noisy Image & Overfitting

- High artifacts and noise on image generations
- Image Background-Foreground almost identical to the instance images



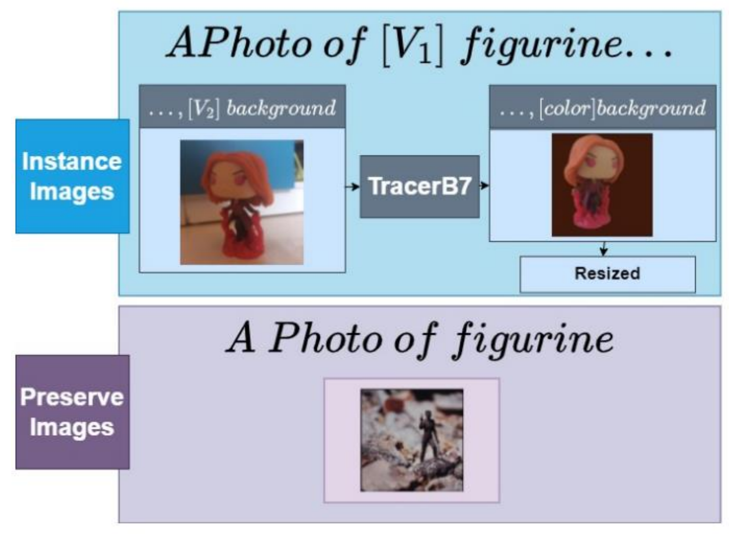

### Foreground-Background Entanglement

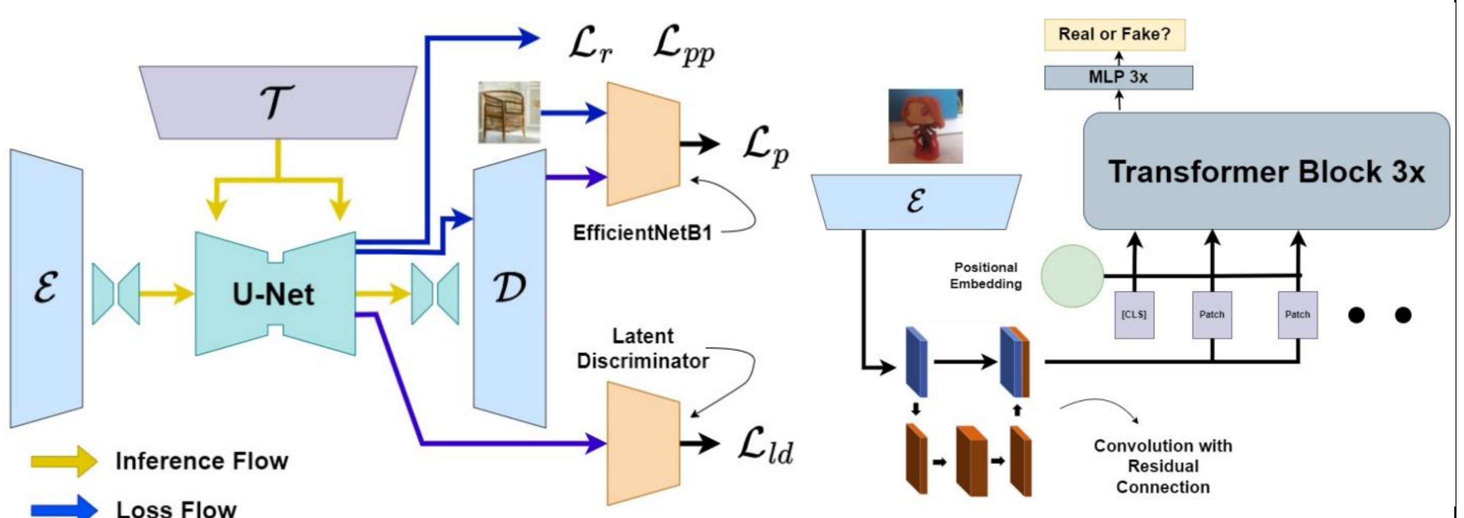
- Background information also learned by the model
- Manipulation towards foreground also applied to the background and vice versa
- Decreasing learning rate decrease both foreground and background alignment towards the prompt and object

## Motivation

- Create reliable inanimate object image generation
- Mimic product photoshoot by utilizing Diffusion Model on both foreground and background. This approach enable broader flexibility compare to just background inpainting
- Promote lightweight fine-tuning technique so that the method is more accessible to a wider range of user.

## Proposed Method



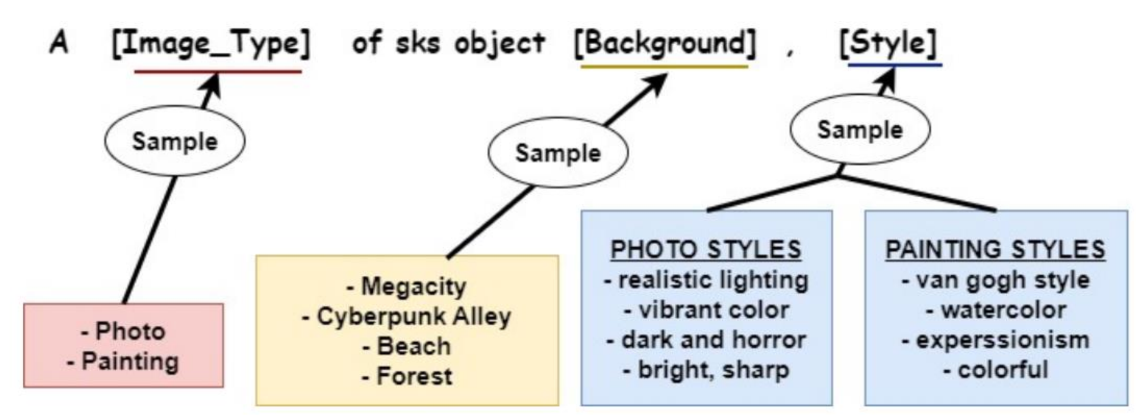

### Image Preparation

- We adopt and extends Dreambooth's training image preparation scheme
- Create variations of provided instance image by changing background color to add variations while retaining lightweightness
- Provide more explicit prompting by separating foreground and background on different clause

### Supervision Mechanisms

- Reconstruction Loss → Rather than using MSE, we use inverse gaussian to calculate deviations between real and predicted noise for instance images
- Prior Preservation Loss → Standard Diffusion MSE loss between the real and predicted noise for preservation images
- Perceptual Loss → Utilizing EfficientNetB1 to calculate perceptual image, this loss only applied to certain amount of steps to prevent overfitting
- Latent Discriminator Loss → We introduce a light transformer based discriminator to discriminate straight on the latent space. This latent discriminator is pre-trained with modified image such as negative color, removed background/foreground. This is done to produce a foreground-focused model

## Proposed Evaluation




- To enable evaluation across different prompts, we propose a novel evaluation mechanism that sample prompt from a pre-defined prompt template. This prompt then can be applied to the existing metrics such as DINO, CLIP-T, CLIP-I, FID, LPIPS, SSIM, and PSNR.
- The prompt part consist of Image type, background information, and style information
- This approach is intended to be used as a complementary insight towards the existing methods

## Qualitative Result


**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

## Quantitative Result

### Invariont Prompt Evaluation

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

| Method | | DINO | CLIP-I | CLIP-T | FID | SSIM | PSNR | LPIPS |
|---|---|---|---|---|---|---|---|---|
| Hypnos (Ours) | ● | **0.7851** | **0.8635** | 0.0094 | 3,6032 | 0.5974 | 11.8504 | 0.3850 |
| | ● | **0.6502** | **0.8015** | 0.0067 | **2.3840** | **0.2225** | **9.4634** | **0.4166** |
| | ● | 0.6589 | 0.8369 | 0.0183 | **5.6330** | 0.3876 | **10.5387** | 0.4624 |
| Dreambooth (LR=1e-6) | ● | 0.6422 | 0.7935 | 0.0183 | **2.9873** | **0.6056** | **12.2883** | **0.3663** |
| | ● | 0.5012 | 0.7404 | **0.0549** | 13.1933 | 0.1645 | 9.0604 | 0.4583 |
| | ● | **0.7130** | **0.8753** | **0.0458** | 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 | **0.4622** |
| Textual Inversion | ● | 0.4934 | 0.6469 | **0.0417** | 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 | **0.3929** | 9.6875 | 0.5160 |

### Varying Prompt Evaluation

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

| Method | | DINO | CLIP-I | CLIP-T | FID | SSIM | PSNR | LPIPS |
|---|---|---|---|---|---|---|---|---|
| Hypnos (Ours) | ● | **0.7070** | 0.7883 | 0.0200 | 11.0675 | 0.5139 | 10.9563 | 0.4626 |
| | ● | **0.5461** | **0.6572** | **0.0326** | **8.6402** | **0.1797** | 8.8435 | 0.5039 |
| | ● | 0.4920 | 0.6462 | 0.0242 | 22.5143 | 0.2863 | **9.7863** | 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 | **0.8028** | 0.0204 | **5.1134** | **0.5589** | **11.6786** | **0.4089** |
| | ● | 0.4656 | 0.6424 | 0.0179 | 17.1525 | 0.1650 | **9.1083** | **0.4583** |
| | ● | **0.5826** | **0.7704** | **0.0336** | **10.4226** | **0.3325** | 9.7451 | **0.4830** |
| Textual Inversion | ● | 0.3131 | 0.4355 | **0.0297** | 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 |

We view quantitative metrics as a supplementary insight rather than an absolute measure of the model overall quality. In some cases lower score is expected, for instance varying prompt often overfitted image scores higher than images that able to align better to the given prompt.

## Conclusion

- **Foreground-Background Disentanglement**

We show an effective approach to enable disentanglement between foreground and background. It is now possible to reliably control scene without subject degradation

- **Clean Image**

Our proposed method capable of creating noiseless image and providing more flexible semantic control through the new hyperparameters

- **Insightful Evaluation**

Varying prompt evaluation opens a new insight along with the existing evaluation methods

## References

1. . 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. 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)