Table 3: Proportion of Latent Discriminator's Training set, the top three correspond to the real images set while the bottom five correspond to the fake images set

Image Type	Proportion	
Instance Image	30%	
Colored Background	18%	50%
Colored Background + resize	2%	
Preservation/Class Image	17,50%	
Negative Foreground (Instance Image)	9,75%	
Masked Foreground (Instance Image)	3,25%	50%
Colored Background + Negative Foreground (Instance Image)	4,88%	
Colored Background + Masked Foreground (Instance Image)	14,63%	

A Preliminary

In this supplementary document, we provide implementation details to ensure reproducibility (written in Appendix B), which includes in-depth strategy and intuition of our approach. Within this section, we include the additional details on the proposed supervision mechanism (Appendix B.1) and configurations on our model architecture (Appendix B.2) according to the main paper's references. Additionally, to elaborate further on our methodology, we present additional visualizations and a detailed discussion of the dataset and the inverse Gaussian function (written in Appendix C and Appendix D). Finally, to offer a comprehensive overview of Hypnos, we include further ablation studies and an analysis on failure cases (written in Appendix E and Appendix F).

B Implementation Details

To train the models we utilize L4 GPU with 22 GB VRAM, however we also had tested that V100 GPU with 16 GB VRAM is sufficient.

All of the models that are used in Sec. 4.1 are trained for 800 steps except for Textual Inversion, which we trained for 1500 steps. We picked 1500 steps for Textual Inversion [6] because it took approximately the same training time as the 800 steps Dreambooth [21]. These models are also trained on 8-bit Adam and quantized VAE. Importantly, to ensure reproducibility, we have set all random seeds to 42. Our code is written in the Pytorch framework.

B.1 Perceptual Loss

As mentioned on Sec. 3.4, to ensure foreground consistency, we opt to set larger weight for the shallow activations, which in detail are **0.35 for the second block activations**, **0.45 for the third block activations**, and **0.2 for the fourth block activations**. Note that this weight number is mostly arbitrary. We advise adjusting the weight to adapt based on the given object and use cases.

Fig. 6: Instance Images, from left to right, top to bottom, Funko figurine, Rattan chair, and Lego Robot.

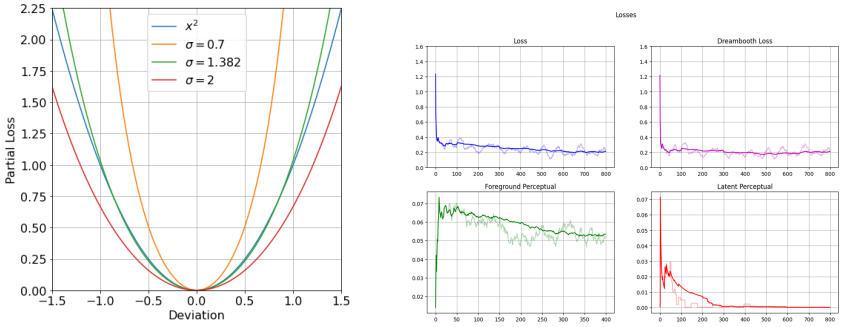
B.2 Latent Discriminator

As seen on Fig. 3a, the images are firstly encoded into latent space using VAE encoder ($\mathcal{E}$) and then fed through a classic inverted bottleneck convolutional layer, which yields a higher level feature extraction. The output then gets concatenated with the original latent image. It then split into 8x8 patches and summed with positional embeddings. The transformer consists of 3 blocks with 3 MLP heads for the [CLS] output embedding.

To prepare the Latent Discriminator with enough knowledge, we pretrained the model for 600 steps. We prepared a dataloader with diverse sampled image types as seen on Tab. 3. These image types are utilized to ensure that the discriminator has a good understanding of foreground distinction, color, and structure. These image types are divided into real images and fake images. The procedure to generate the real images is identical to the one used on the main training loop as mentioned on Sec. 3.3. On the other hand, the fake image extraction procedure introduces a more diverse approach; the most intuitive method is to include preservation images as the fake samples; this addition is meant to enforce the model's understanding of structures and differentiate one object from another. Besides using the preservation images, we also used the altered foreground version of the instance images. The first modification is the usage of negative colors with the same background; this approach ensures the latent discriminator understands color consistency. Second, the masked foreground strategy is applied to teach the model to avoid relying on the background region to distinguish real or fake images. Note that we aim to ensure the discriminator has a clear focus on the foreground region. Hence, we opt to label images with foreground and background modification as fake and real data, respectively.

B.3 Evaluation

To evaluate metrics on both Prompt Invariant and Varying, we calculate the mean and standard deviations of each metric across 50 generated images. For each image that is generated by Hypnos, we compared it with each instance-images, as shown in Fig. 6. For example, in a dataset that is comprised of 4



(a) Inverse Gaussian comparisons **(b)** Learning Curve

Fig. 7: (a) shows varying σ affects Inverse Gaussian function $((1/gaussian(x)) - 1)$ in partial loss space compared to quadratic MSE function, (b) shows the stability of learning curve.

images, by this approach, the metrics will be calculated across 200 different evaluations (4×50 evaluations).

C Dataset Images

Fig. 6 shows the instance images that we utilized for evaluation purposes on Sec. 4.1. In the case of Funko figurine and Lego Robot, we use personal images taken using mobile phones to show that using amateur photos is already sufficient to make Hypnos generate visually pleasing image quality. Note that poor lighting and shadows, as seen in Fig. 6, might still cause color distortions.

D Inverse Gaussian

As seen in fig Fig. 7a, employing Eq. (2) allows for the flexible adjustment of the steepness of the curve. This loss is an exponential function; hence, it exhibits a far larger loss for high deviations compared to MSE. Usually, this approach is avoided because of the possible training instability. Fortunately, Hypnos finetunes a pretrained network; therefore, the loss is often already very close to 0 and rarely exceeds 1, resulting in a stable training cycle as seen on the learning curve (Fig. 7b).

E Supplementary Ablation Study

E.1 Variation of Generated Images

Throughout our observations throughout the evaluation process on both qualitative and quantitative metrics, we observed that Dreambooth family techniques, which include our proposed method, have high variations in generated images.

Fig. 8: Examples of Failure Cases. (left) shows an image that fails to adapt to the given prompt, (right) shows an altered object with one arm missing

We observed that even with the same trained model, by reevaluating the evaluation prochedure, it is possible to acquire a score more than one standard deviation away from the previous measurement. Qualitatively, Dreambooth often generates a high noised image, and on the other generated images, it is possible to stumble across neat images output as shown in Fig. 9, Fig. 10, and Fig. 11. On the other hand, Hypnos exhibit a more consistent image quality throughout the generation process. Based on this understanding, we recommend sampling several images and conducting a manual assessment prior to their utilization in downstream tasks.

E.2 Perceptual Loss and Latent Discriminator Loss

Fig. 12 shows the visualization of ablation study experiments mentioned on Sec. 4.2. As shown in figure Fig. 12, Perceptual Loss and Latent Discriminator are complementary losses. In other words, the absence of one results in a deterioration of the quality of the generated images. This is reasonable as both loss works in different spaces and perspectives as explained on Sec. 3.4.

Both losses are also unsuitable to replace reconstruction loss to guide the optimization process (shown by the increase of weight on the first and fourth columns of Fig. 12). This is supported by the solid mathematical derivation of reconstruction loss and the nature of Perceptual Loss and Latent Discriminator Loss that may vary based on the reliability of the corresponding neural network models (e.g., EfficientNetB1, ViT).

F Failure Cases Analysis

Throughout our evaluation, we also assessed some failure cases to better understand the behavior and limitations of Hypnos. Fig. 8 shows examples of failure

cases generated by Hypnos. The first shows a blank background when given a complex prompt. We suspect this is partly due to the base model limitation, we justify this by confirming that the base model still fails to generate the same prompt even with a normal chair. We also observed that Hypnos tends to be more conservative in generating foreground variations with a trade off in foreground consistency. This phenomenon can be minimized by decreasing the changed background ratio and decreasing the perceptual and latent discriminator strength.

We observed that the provided preservation dataset also influenced background and scene generation capability. Supplying several preservation images with a specific background can enable the model to generate similar backgrounds. We anticipate that this effect may be less significant for larger base models, as they typically possess broader text-to-image (T2I) capabilities. However, further justification and experiments on this matter are beyond the scope of this work and may be addressed in future research.

The second image from the left (Fig. 8) showed a less similar foreground with the absence of one forearm. This is due to the limitation of the background remover model, which accidentally removes the arm. This issue can be overcome by excluding the problematic image, reducing the change in background ratio, or even replacing the background remover model. On the other hand, the underfitting can be explained by the complexity of the object. Hence, this can be trivially overcome by increasing the learning rate. Hypnos shows to be compatible with a high learning rate as it is still able to generate noiseless images.

In other cases where generating a better foreground is more favorable than training efficiency, we advise incorporating a stronger and deeper discriminator as it may increase the image quality. Despite the limitations above, Hypnos proves that the proposed finetuning strategy can accommodate the current prominent T2I model in handling foreground-focused generative task in a straightforward manner.

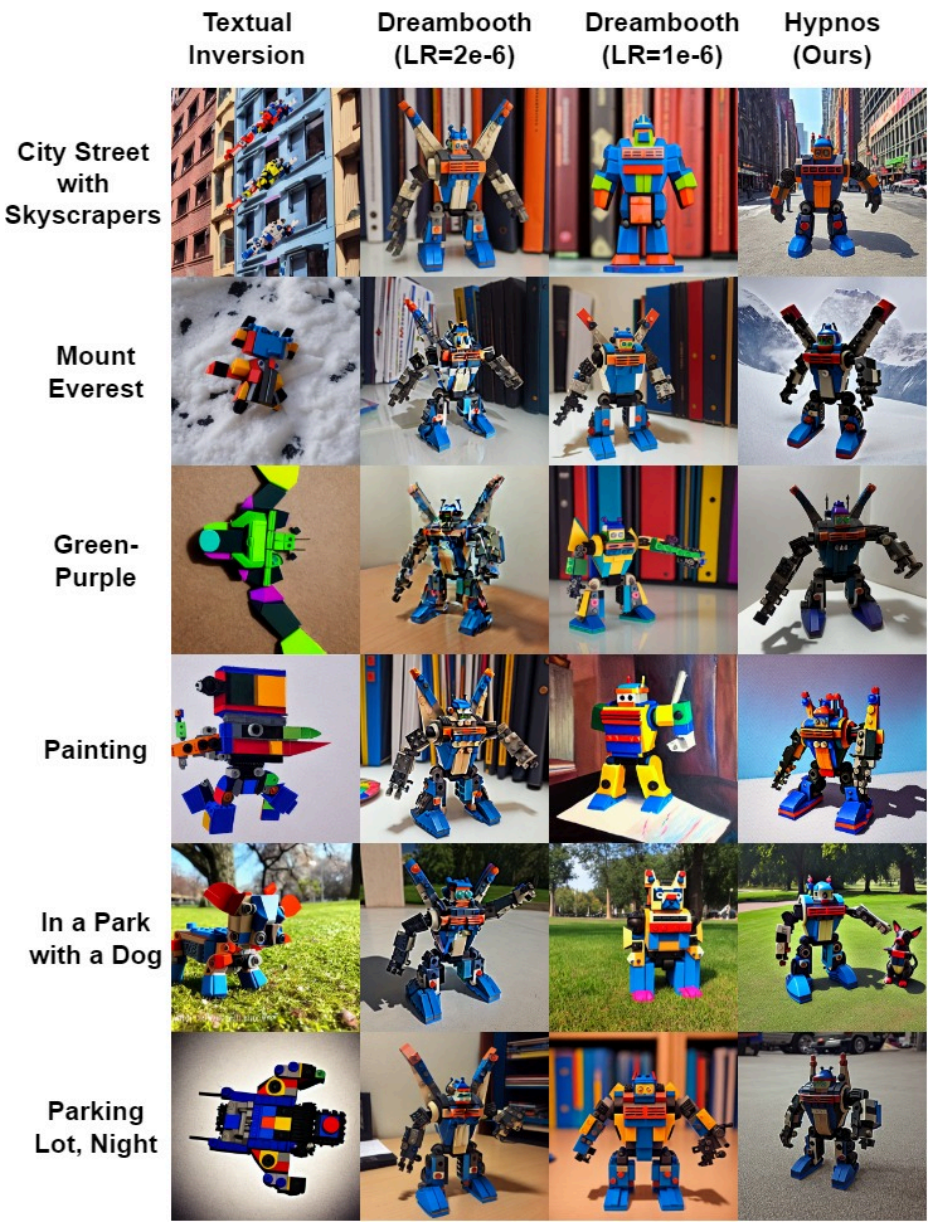


Fig. 9: Lego Robot comparison (electronic screen viewing is advised)

Fig. 10: Funko Figurine comparison (electronic screen viewing is advised)

Fig. 11: Rattan Chair comparison (electronic screen viewing is advised)

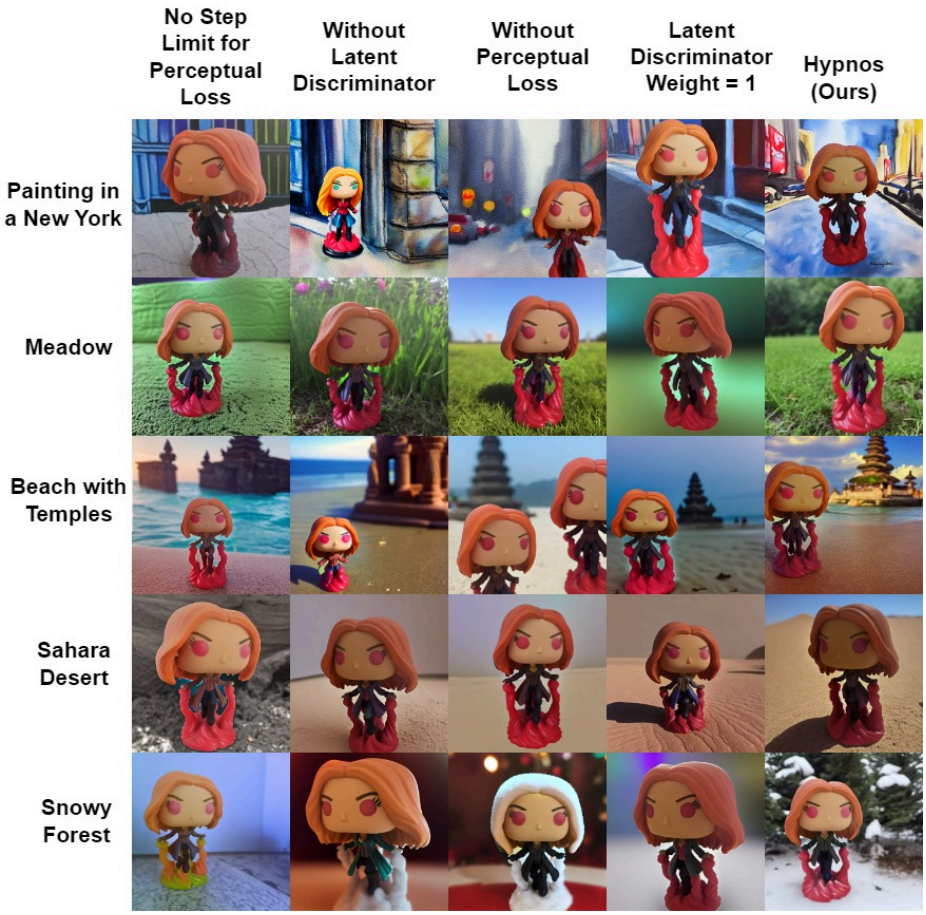


Fig. 12: Funko Figurine ablation study (electronic screen viewing is advised)