NeurIPS Rebuttal

We thank all the reviewers for the hard work. We want to highlight a few things:

Can SuTI be combined with DreamBooth?

One major clarification we want to make is that our model is not exclusive to DreamBooth [6]. Instead, SuTI can be combined with DreamBooth [6] naturally to achieve better results. Specifically, given K reference images regarding a subject, we can randomly feed 1 image as the condition and another random image as the target output. By training for 500 steps like DreamBooth [6], we can evaluate our Dream-SuTI model on the given subject. In this case, because the model already knows the subject, we only need to give the model on demonstration, which can lower the inference to roughly the same level as DreamBooth itself. We show an example in Figure 13, where we pick

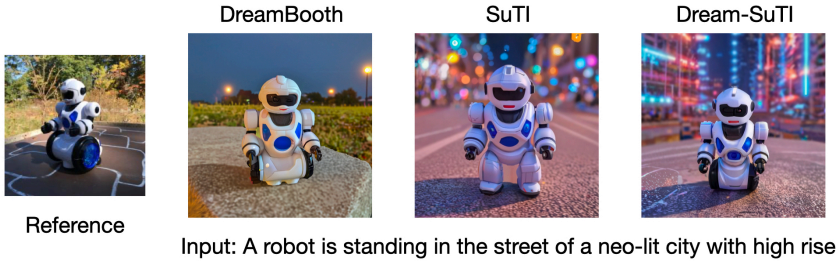

Figure 13: Comaprison between DreamBooth, SuTI and Dream-SuTI.

a failure example from SuTI to investigate whether Dream-SuTI can improve it. We found that DreamBooth does not have strong text alignment, while SuTI's subject lacks fidelity (the robot uses wheel instead of legs). By doing DreamBooth enhancement, Dream-SuTI is able to generate images not only faithful to the subject but also to the text description. We further conduct quantitative human evaluation on Dream-SuTI and report the results below: We can observe that Dream-SuTI can

Methods	Inference Time	Subject	Text	Photorealism	Overall
DreamBooth	10 secs	0.88	0.82	0.98	0.77
SuTI	30 secs	0.90	0.90	0.92	0.82
Dream-SuTI	15 secs	0.92	0.92	0.94	0.87

Table 3: Quantitative Human evaluation on Dream-SuTI

further improve the overall score from 0.82 to 0.87, yielding 5% improvement over SuTI, and 10% improvement over DreamBooth.

Why do we need to use DreamBooth to synthesize target output?

An important ablation study we perform is to use the image clusters alone to train SuTI without any DreamBooth experts. Specifically, for each cluster with k images, we adopt $k-1$ as the demonstration to train SuTI to generate the k-th image as the target. This approach can lower the computation cost significantly. However, our attempt shows that SuTI will simply learn to copy-paste from the demonstration due to the similarity of images within the cluster. We demonstrate some examples

Figure 14: Visualization of the clustered images.

in Figure 14, the images within the cluster are almost replicate of each other, which causes SuTI to fall into a local optimum to do copy-paste. The benefit of DreamBooth is that we can control the diversity by feeding in highly versatile prompts.