

---

# Supplementary Material: Emergent Properties of Efficient Fine-Tuning in Text-to-Image Models

---

Komal Kumar<sup>1\*</sup> Rao Muhammad Anwer<sup>1</sup> Fahad Shahbaz Khan<sup>1</sup>  
Salman Khan<sup>1</sup> Ivan Laptev<sup>1</sup> Hisham Cholakkal<sup>1</sup>

<sup>1</sup>Mohamed bin Zayed University of Artificial Intelligence Abu Dhabi, UAE  
{komal.kumar, rao.anwer, fahad.khan,  
salman.khan, ivan.laptev, hisham.cholakkal}@mbzuai.ac.ae

🌐 **Dream Branch Plus Comparison:** <https://anonymousdreambranchplus.netlify.app>

🔗 **Omnigen-Visualcloze:** <https://anonymouscloze.netlify.app/>

🔗 **InsT Objects Qualitative Results:** <https://anonymousinstobjets.netlify.app>

## S1 Further Quantitative and qualitative results

### S1.0.1 Image generation consistency

In this section, we present a detailed comparison of various fine-tuning methods for different subjects using the CLIP-image score on the Dreambooth dataset, as shown in Table 1. The methods include the proposed DEFT, PaRa [1], and LoRA [2], alongside previous approaches such as Texture Inversion and DreamBooth. The table compares performance across three distinct subject categories: BEAR\_PLUSHIE, CAT, and DOG8. Our proposed DEFT method, with a rank of 4, achieves high CLIP-image scores of 0.8339 for BEAR\_PLUSHIE, 0.9280 for CAT, and 0.8721 for DOG8. Notably, PaRa with rank 4 shows slightly improved results, especially for DOG8, with a score of 0.8838. In contrast, LoRA methods, particularly with ranks 4 and 8, show lower performance scores, especially for BEAR\_PLUSHIE. Compared to previous methods such as PaRa [1] and SVDIFF [4], our proposed methods, DEFT, exhibit competitive or superior results in terms of image-text alignment across all subject categories, underlining their effectiveness for multimodal image generation tasks.

"A photo of [V]"	BEAR_PLUSHIE	CAT	DOG8
DEFT (rank=8) (Our)	<b>0.8415</b>	<b>0.9504</b>	0.8882
DEFT (rank=4) (Our)	0.8339	0.9280	0.8721
PaRa [1](rank=4)	0.8271	0.9315	0.8780
PaRa [2] (rank = 8)	0.8051	0.9467	<b>0.8955</b>
LoRA [2](rank=4)	0.7741	0.8057	0.7773
LoRA [7] (rank = 8)	0.7943	0.8583	0.8295
SVDIFF [4]	0.7818	0.8854	0.8363
DREAMBOOTH [7]	0.7921	0.8893	0.8392
TEXTURE INVERSION [3]	0.7421	0.8048	0.7432

Table 1: Comparison of Various Methods for Different Subjects Using Clip-Image Score on Dreambooth Dataset: The table presents the comparison of different fine-tuning methods on the Dreambooth dataset, evaluated using the CLIP-image score. It highlights the performance of the proposed DEFT, PaRa, and LoRA methods against previous approaches, including Texture Inversion and DreamBooth, across multiple subject categories like BEAR\_PLUSHIE, CAT, and DOG8.

---

\*Corresponding author.

### S1.0.2 Qualitative comparison

Furthermore, the Figure 1 presents qualitative results comparing different fine-tuning strategies and DEFT applied to the Unified Omnigen model. It showcases various image generations of a dog across different environments and prompts, including a lush green field, a beach, a snowy landscape, a cityscape, a garden, and a forest. Each model—Base, LoRA, PaRa, and DEFT—produces distinct results, emphasizing how these fine-tuning methods affect image generation based on specific prompts. The outcomes demonstrate the ability of these techniques to enhance the generalization and adaptivity of the model while maintaining high-quality, realistic results. The comparisons underline the impact of efficient fine-tuning in improving the model’s ability to generate diverse, accurate images across various scenarios.



Figure 1: Qualitative Results on Unified Omnigen Model Comparing Efficient Fine-Tuning and DEFT: This figure presents qualitative results comparing efficient fine-tuning strategies and DEFT on the Unified Omnigen model. The outcomes demonstrate the capability of these techniques to enhance model generalization and adaptivity while maintaining high-quality results.

Furthermore, the Figure 2, demonstrates the model’s impressive ability to generalize across a wide range of unseen prompts. The image features four different outputs generated based on distinct themes: abstract, fantasy, futuristic, and historical prompts. Despite the varied nature of the inputs, the model consistently produces high-quality results, showing its adaptability to different styles and concepts. The figure emphasizes the model’s versatility, highlighting its capacity to maintain visual coherence and output quality across diverse scenarios, from abstract landscapes to historical depictions. This illustrates the robustness of the model in handling various types of prompts while ensuring consistency in the final image outputs.



Unseen prompt abstract    Unseen prompt fantasy    Unseen prompt futuristic    Unseen prompt historical

Figure 2: Diverse Prompt Generalization: This figure shows the generalization capabilities of the model across diverse prompts, emphasizing its ability to handle a variety of inputs while maintaining consistent output quality.

### S1.0.3 Qualitative and quantitative differences

The qualitative and quantitative comparison between the methods DEFT and LoRA, as shown in both the table 2 and the images 3 on dreambench plus [5] with SDXL [6] finetuning, reveals distinct strengths for each model in generating images of cats and horses. From the quantitative perspective, LoRA consistently achieves higher scores across DINOv1 and DINOv2 for both the Kitten and HORSE images. For example, LoRA outperforms DEFT in DINOv1 and DINOv2 for the Kitten image (83.5538 vs. 79.9416 for DINOv1, and 72.2653 vs. 65.0358 for DINOv2), and similarly for the HORSE images (83.5538 vs. 77.8501 for DINOv1, and 72.2653 vs. 65.0358 for DINOv2). These results suggest that LoRA is better at capturing complex features and achieving higher-quality representations in these metrics, which might reflect its greater flexibility and artistic adaptability.

In contrast, DEFT demonstrates a stronger performance in CLIP-I and CLIP-T for some images, especially for the Kitten images (83.5867 for DEFT vs. 81.3446 for LoRA in CLIP-I), indicating its ability to produce more realistic, detailed representations that preserve the original essence of the animals. DEFT’s output tends to have clearer textures, sharper details, and more lifelike features, showcasing its strength in realism and faithful reproduction.

The images generated by DEFT are more grounded in reality, with clear textures and natural settings. In the case of the HORSE images, DEFT retains more authentic anatomical features and textures, reflecting a more realistic depiction of the animals. On the other hand, LoRA brings a more artistic flair to the HORSE images, with creative use of colors, dynamic poses, and abstract elements. While LoRA’s outputs are more vibrant and imaginative, they may not always preserve the natural look and feel of the animals as consistently as DEFT does.

Despite LoRA’s higher scores in DINOv1 and DINOv2, these results do not fully capture its ability to maintain realistic features across all images. LoRA excels in producing creative, artistic representations, but at the cost of some consistency in realism. DEFT, with its emphasis on realism, demonstrates more stable, high-fidelity outputs, particularly for complex subjects like horses.

This analysis shows that while LoRA excels in artistic flexibility and creative interpretation, achieving higher DINOv1 and DINOv2 scores, DEFT remains superior in generating more realistic and detailed images. The choice between the two methods ultimately depends on the desired outcome—whether the goal is to prioritize artistic creativity or to maintain realistic accuracy.

### S1.0.4 Scene personalization

Figures 4 and 5 showcase the model’s scene personalization capabilities, demonstrating its proficiency in generating high-quality visual content with specific environmental characteristics.

Figure 4, titled Church Rock Scene Personalization, illustrates how the model adapts the Church Rock scene to various settings. These scenes include dynamic backgrounds, such as a pool surrounded by palm trees, a futuristic city at night, a snowy mountain top, a crowded street market, and a forest with autumn leaves. Each personalized scene is a result of fine-tuning, reflecting how the model can generate diverse visual representations of the same object in unique environments, showcasing its flexibility in handling scene-specific details.

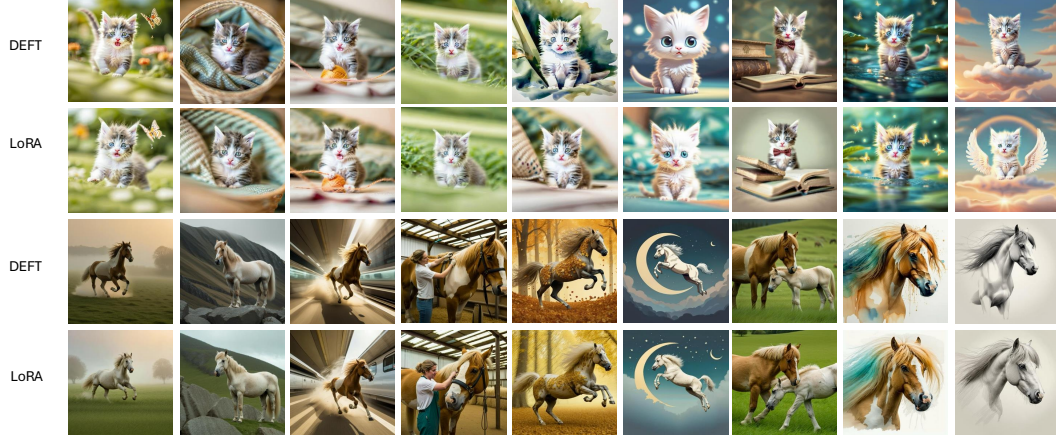


Figure 3: Qualitative comparison of the DEFT and LoRA methods for generating images of cats and horses. The first row shows images of cats, while the second row shows horses. DEFT produces realistic, detailed images, while LoRA introduces more artistic and stylized elements, showcasing its flexibility in adapting to creative representations.

Method	Image	DINOv1	DINOv2	CLIP-I	CLIP-T
DEFT	Kitten	79.9416	73.5553	83.5867	35.5409
DEFT	Stork	67.9085	62.7926	76.3276	36.8799
DEFT	Kitten 2	77.8501	65.0358	77.6003	36.7579
DEFT	HORSE	77.8501	65.0358	77.6003	36.7579
LoRA	Kitten	82.9492	77.9071	86.7823	33.7977
LoRA	Stork	70.1220	61.8301	74.9231	36.9636
LoRA	Kitten 2	83.5538	72.2653	81.3446	34.1923
LoRA	HORSE	83.5538	72.2653	81.3446	34.1923

Table 2: Comparison of DEFT and LoRA across multiple evaluation metrics (DINOv1, DINOv2, CLIP-I, CLIP-T) for images of cats and horses. The table highlights how LoRA consistently achieves higher scores in DINOv1 and DINOv2, indicating its strength in capturing complex features, while DEFT excels in CLIP-I and CLIP-T, reflecting its focus on realism and detailed preservation of the original subjects.

Figure 5, titled Table Scene Personalization, further exemplifies the model’s ability to adapt to specific environments. In this case, the model personalizes a simple table scene by generating various configurations of objects like bottles and caps, based on detailed prompts. The generated scenes show different bottles, one filled with orange liquid and another empty, both with distinct cap colors. This reflects the model’s ability to generate high-quality content by adapting to specific setups, maintaining both object clarity and spatial coherence within the scene.



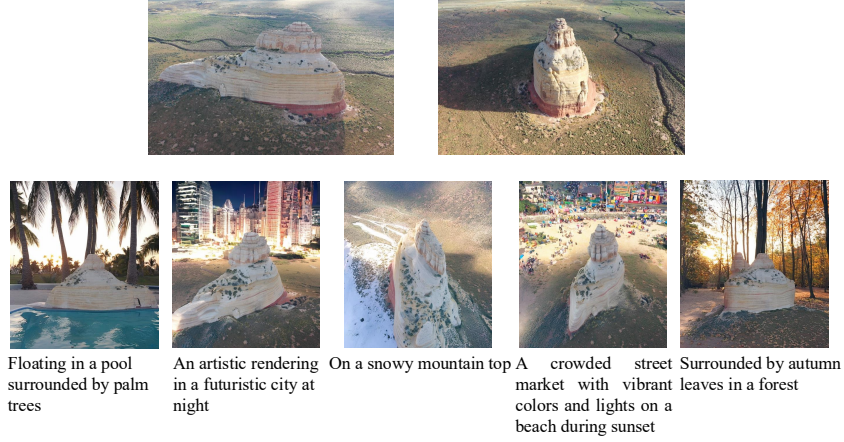


Figure 4: This figure illustrates the scene personalization capabilities of the model using the church rock scene, showcasing how fine-tuning allows for detailed control over scene-specific characteristics.

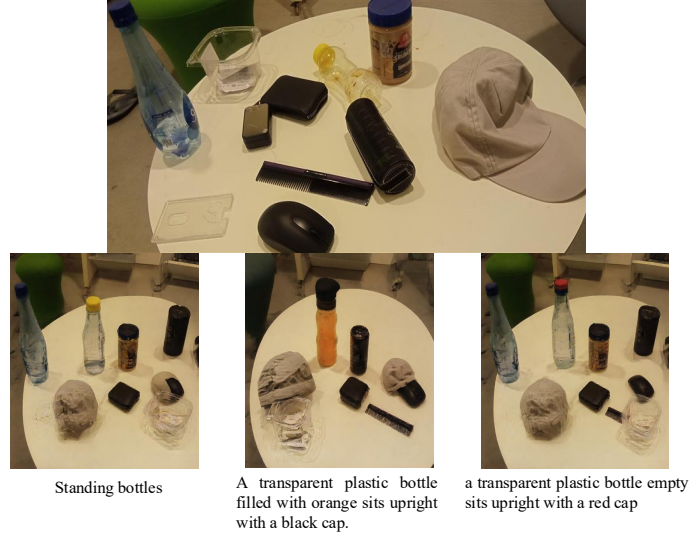


Figure 5: Table Scene Personalization: The figure demonstrates how the model personalizes a table scene, reflecting the ability to adapt and generate high-quality visual content in specific environments.

## References

- [1] Shangyu Chen, Zizheng Pan, Jianfei Cai, and Dinh Phung. Para: Personalizing text-to-image diffusion via parameter rank reduction. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [2] cloneofsim. Low-rank adaptation for fast text-to-image diffusion fine-tuning, 2022.
- [3] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.
- [4] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdif: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7323–7334, 2023.
- [5] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

- [7] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dream-booth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.