Qualcomm
AI research

Hollowed Net for On-device Personalization of Text-to-Image Diffusion Models

Wonguk Cho[1], Seokeon Choi[1], Debasmit Das[2], Matthias Reisser[2], Taesup Kim[3], Sungrack Yun[1], Fatih Porikli[2]

[1]Qualcomm Korea YH
[2]Qualcomm Technologies, Inc.
[3]Seoul National University

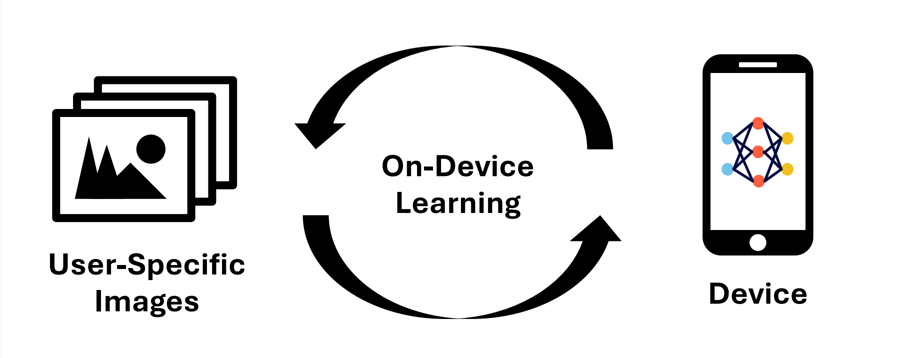
Presenter: Wonguk Cho

Interim Engineering Intern

On-Device Personalization of T2I Diffusion Models

Task Definition

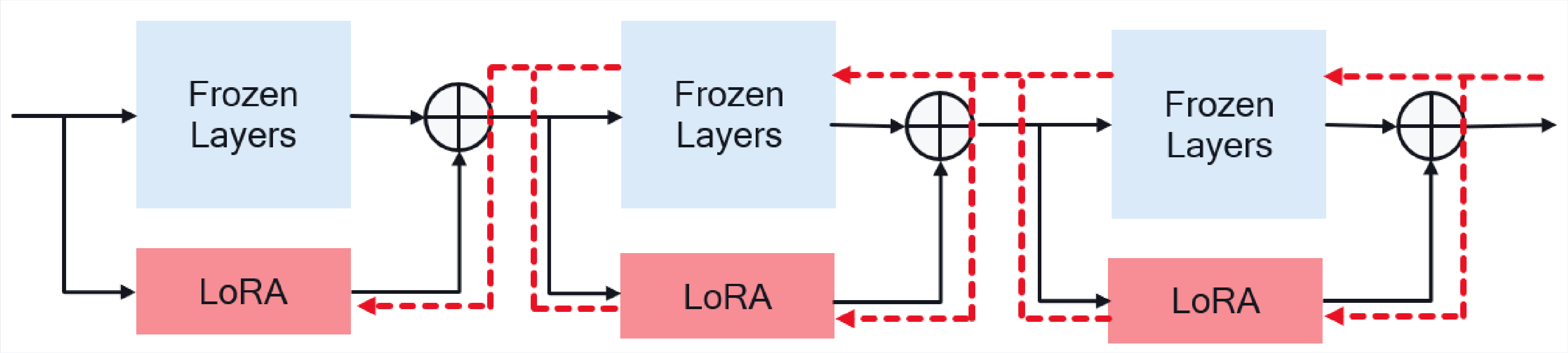
- We aim to enable **on-device personalization** of T2I diffusion models, by **fine-tuning** models on the mobile devices **with user-specific images** for customized generation.
- On-device personalization can make the entire user experience **all-the-more personal** and help **protect users' privacy** since personal information remain solely on the device.



On-Device Personalization of T2I Diffusion Models

Challenges

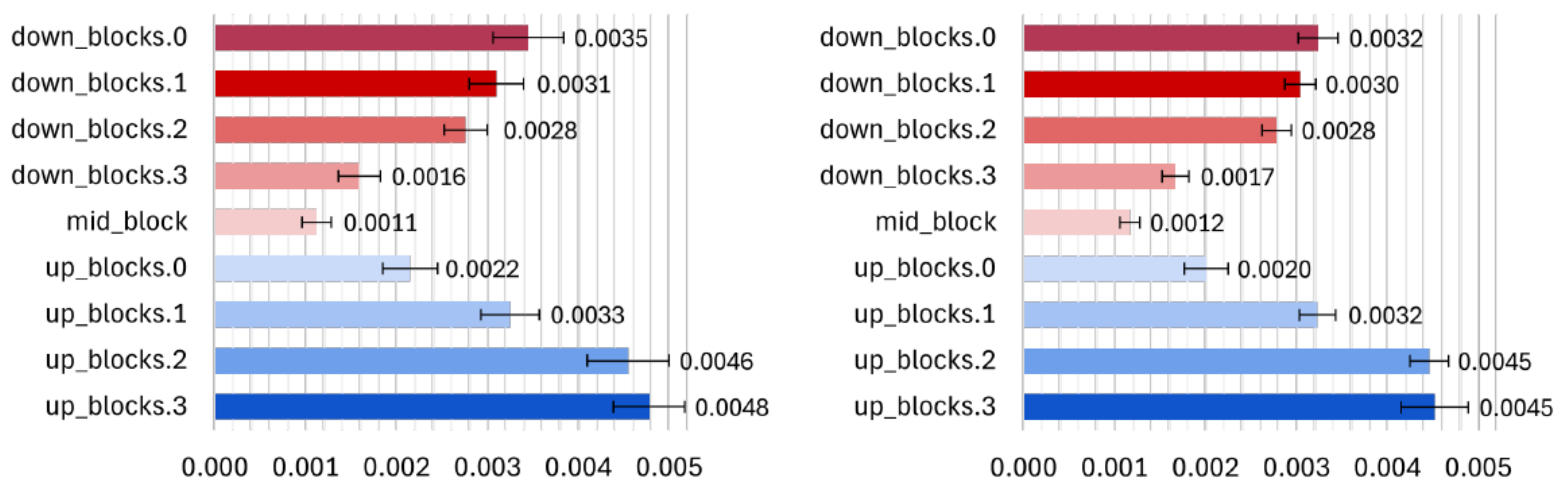
- The challenge of on-device learning stems from **limited computational resources** of end device, particularly in terms of **memory I/O**.
- Existing PEFT methods are **limited in extremely low memory resources** as they require backpropagation over large diffusion models and **do not reduce memory usage from loading model weights.**



LoRA personalization with Hollowed Net

Motivation

- We find that **the blocks around the center** of the diffusion U-Net are **less involved in the personalization**.
- Building on this insight, we design Hollowed Net, which can **temporarily excluding these less significant layers during fine-tuning** to reduce peak memory usage.

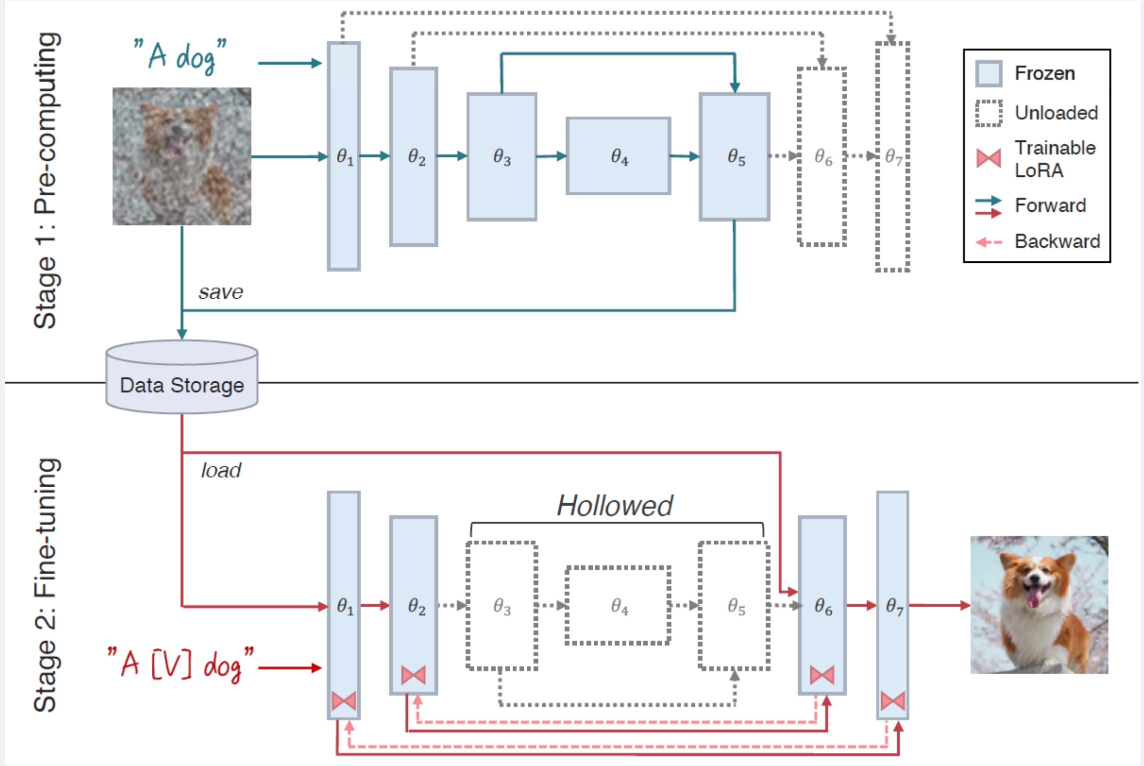


(a) DreamBooth (b) CustomConcept101

LoRA personalization with Hollowed Net

Training w/ Hollowed Net

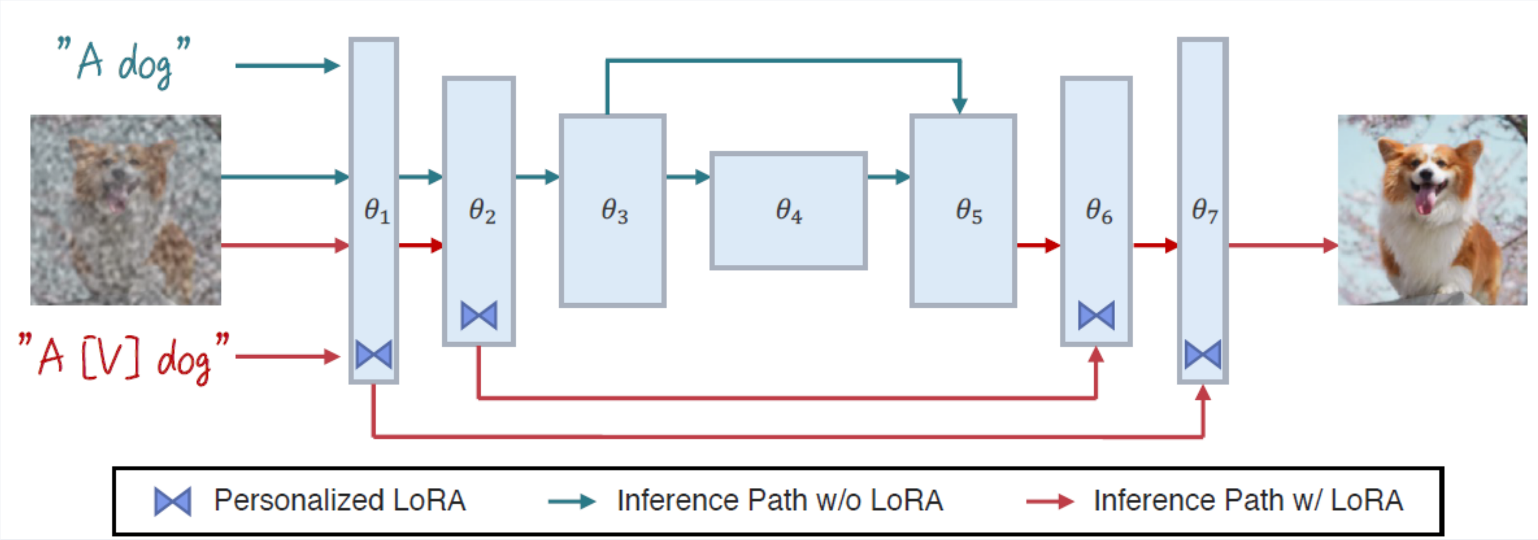
- We propose **two-stage fine-tuning strategy**:
  1. **Pre-computing** intermediate activations of the original diffusion U-Net
  2. **Fine-tuning** the Hollowed Net using the pre-computed activations



LoRA personalization with Hollowed Net

Inference w/ Personalized LoRA

- Hollowed Net **does not need to be held during inference**, by transferring personalized LoRA parameters to the original U-Net.
- We sequentially execute two inference paths, respectively corresponding to each stage of fine-tuning.



Experimental Results

Comparison with Full / LoRA FT

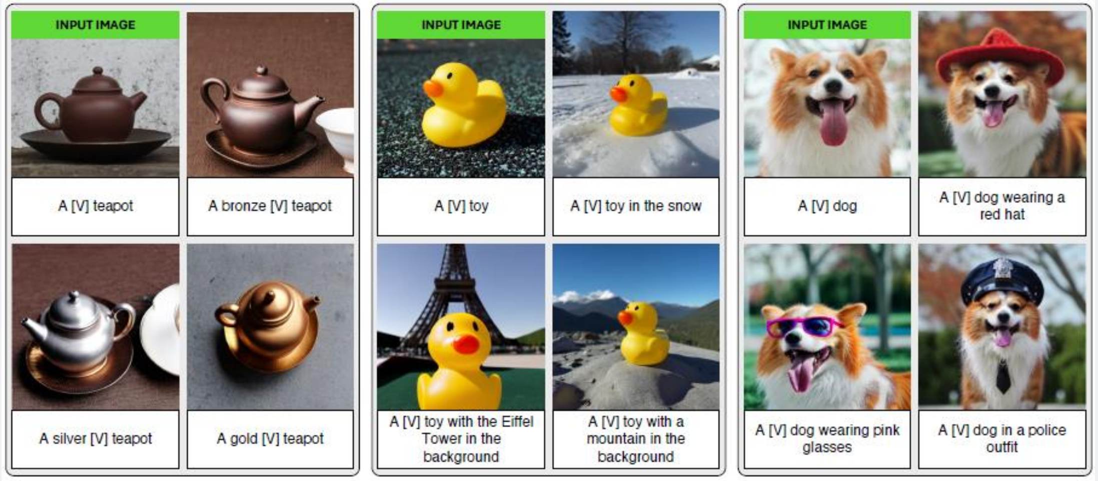
- We conduct experiments with 131 subjects and demonstrate that Hollowed Net achieves high-fidelity personalization results comparable to Full FT while requiring 77% (12.74GB) less GPU memory.
- The required memory for Hollowed Net fine-tuning is only 11% (390MB) more than needed for inference.

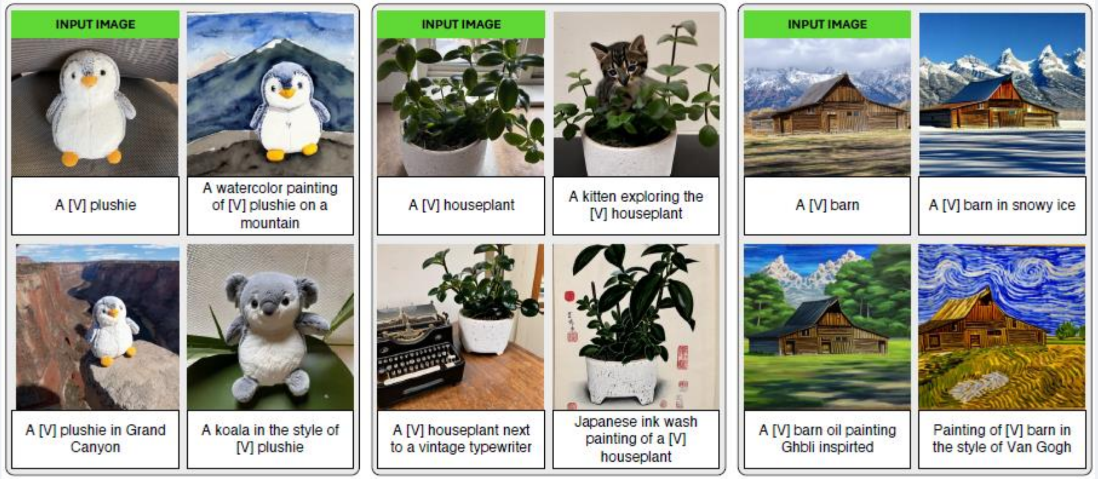
Method	# of Parameters		Training Memory		DreamBooth			CustomConcept101		
	Base	LoRA	Peak	Comp. w/ Inf.	DINO	CLIP-I	CLIP-T	DINO	CLIP-I	CLIP-T
Full FT	866M	-	16.62GB	+376%	0.663 $\pm$0.013	0.802 $\pm$0.007	0.302 $\pm$0.002	0.605 $\pm$0.005	0.773 $\pm$0.006	0.302 $\pm$0.002
LoRA FT (r=128)	866M	27M	5.23GB	+50%	0.658 $\pm$0.001	0.806 $\pm$0.005	0.299 $\pm$0.002	0.603 $\pm$0.008	0.773 $\pm$0.005	0.302 $\pm$0.002
LoRA FT (r=1)	866M	207K	4.84GB	+39%	0.516 $\pm$0.011	0.738 $\pm$0.003	0.314 $\pm$0.001	0.522 $\pm$0.008	0.737 $\pm$0.005	0.305 $\pm$0.001
Hollowed Net (Ours)	**527M**	**24M**	**3.88GB**	**+11%**	0.660 $\pm$0.011	0.805 $\pm$0.006	0.300 $\pm$0.001	0.603 $\pm$0.007	0.773 $\pm$0.005	0.302 $\pm$0.002

Experimental Results

Qualitative Results

- Hollowed Net effectively captures the visual details of the target subjects, while maintaining high text-image alignment for different types of applications including property modification, recontextualization, accessorization, and artistic rendition.

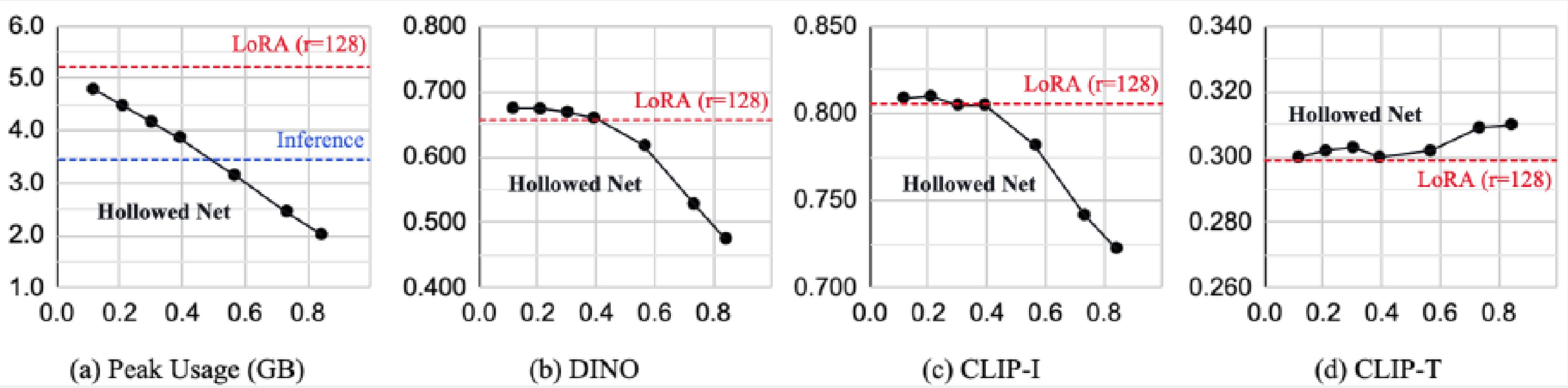




Experimental Results

Fractions of Hollowed Layers

- We find that the model's capacity to preserve subject fidelity remains comparable to or slightly better than LoRA FT **until around 39.2% of layers are hollowed.**
- Users can adjust the fraction of hollowed layers to control the trade-offs between performance and memory requirements, depending on the target application and resources.



Thank you

Nothing in these materials is an offer to sell any of the components or devices referenced herein.

References in this presentation to “Qualcomm” may mean Qualcomm Incorporated, Qualcomm Technologies, Inc., and/or other subsidiaries or business units within the Qualcomm corporate structure, as applicable. Qualcomm Incorporated includes our licensing business, QTL, and the vast majority of our patent portfolio. Qualcomm Technologies, Inc., a subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of our engineering, research and development functions, and substantially all of our products and services businesses, including our QCT semiconductor business.

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries.
Qualcomm patented technologies are licensed by Qualcomm Incorporated.

Follow us on:

For more information, visit us at qualcomm.com & qualcomm.com/blog