Mining your Own Secrets: Diffusion Classifier Scores for Continual Personalization of Text-to-Image Diffusion Models

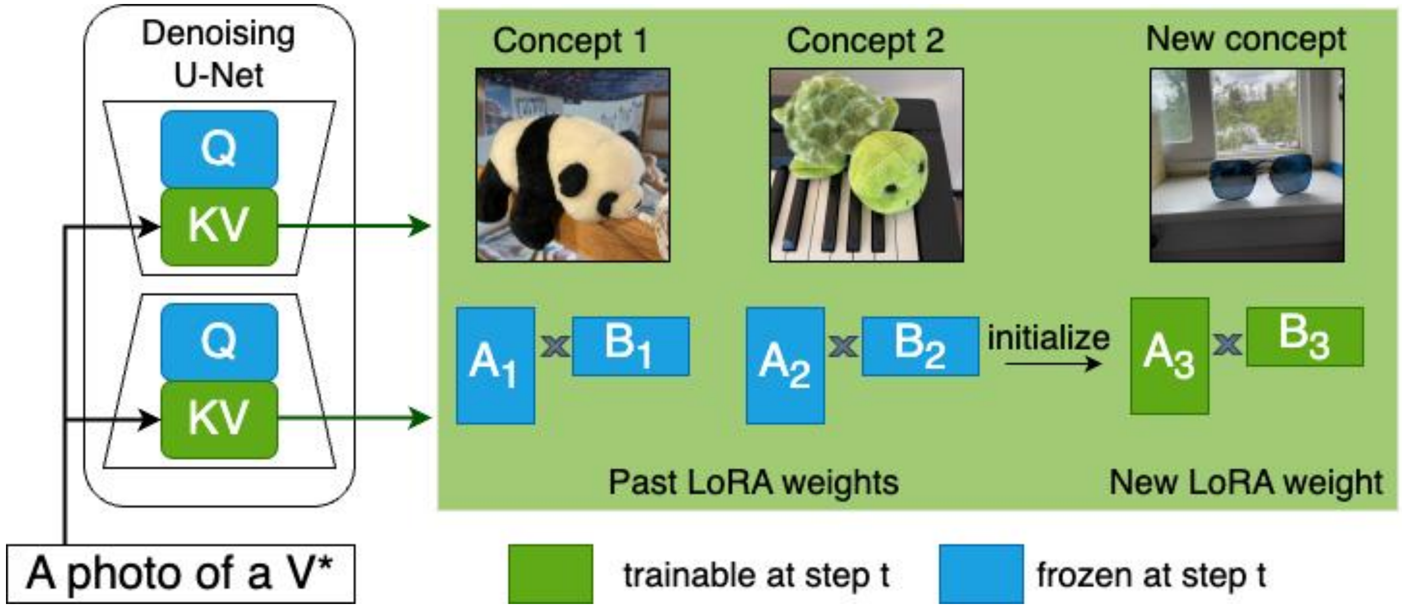
Saurav Jha*, Shiqi Yang, Masato Ishii, Mengjie Zhao, Christian Simon, Muhammad Jehanzeb Mirza, Dong Gong, Lina Yao, Shusuke Takahashi, Yuki Mitsufuji

* Work done as an intern at Sony Japan.

https://srvcodes.github.io/continual_personalization/

Continual Personalization with LoRA

- We acquire one new concept at a time step ‘t’
- We finetune LoRA for K,V layers in the U-Net cross-attention
- Sequential initialization of LoRA layers



Motivation: Class-incremental learning

- Regularization-based methods widely use class-specific information
- E.g. EWC using cross-entropy loss for Fisher Information estimation

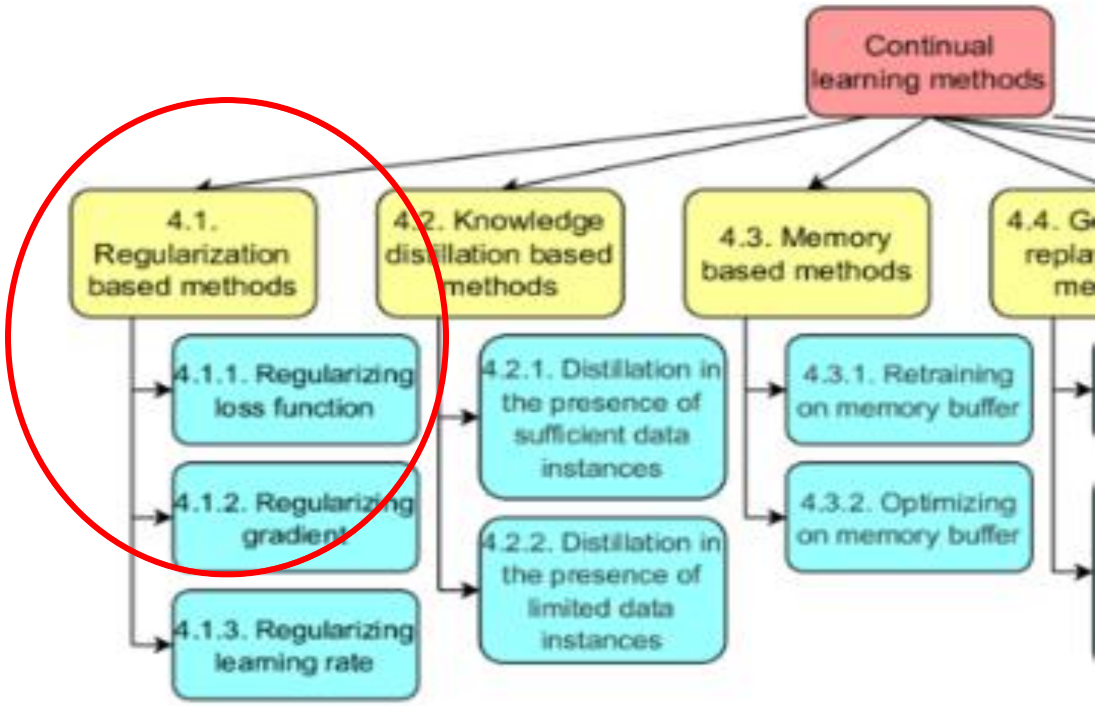


Image source: Qu et al. "Recent Advances of Continual Learning in Computer Vision: An Overview"

Class-specific information in diffusion models

- **Diffusion Classifier (DC) scores:** the softmax over the negative denoising scores w.r.t. each concept [1]

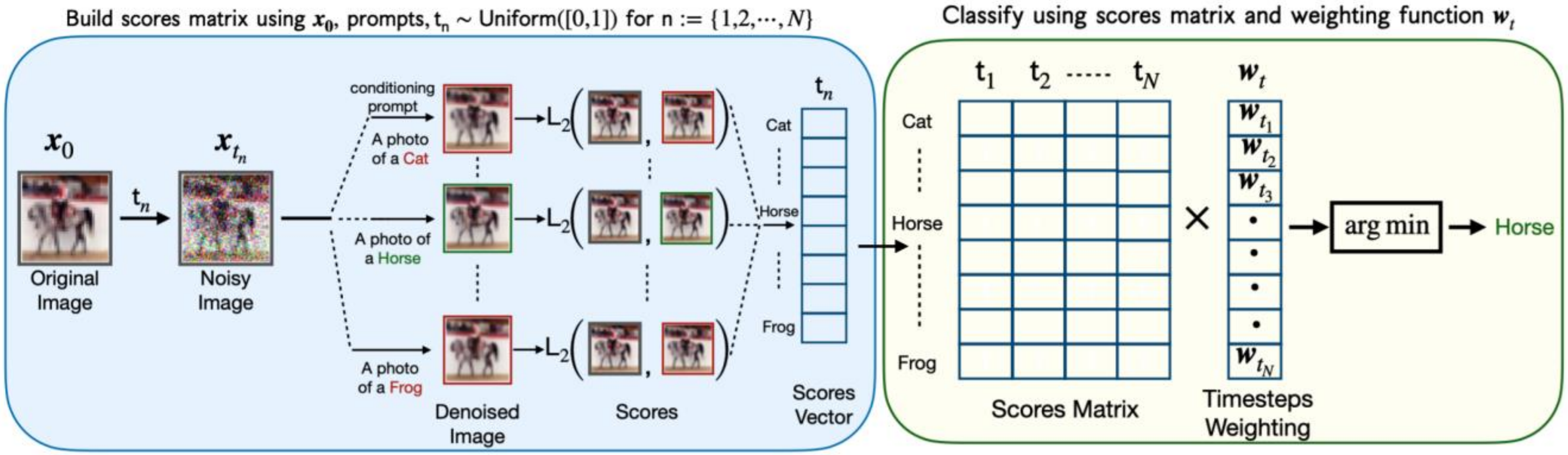


Image source: [1]

[1] Clark, Kevin and Priyank Jaini. "Text-to-Image Diffusion Models are Zero-Shot Classifiers." *NeurIPS 2023*.

Roadmap

Parameter-space regularization framework

Parameter-space reg. with DC scores

Functional-space regularization framework

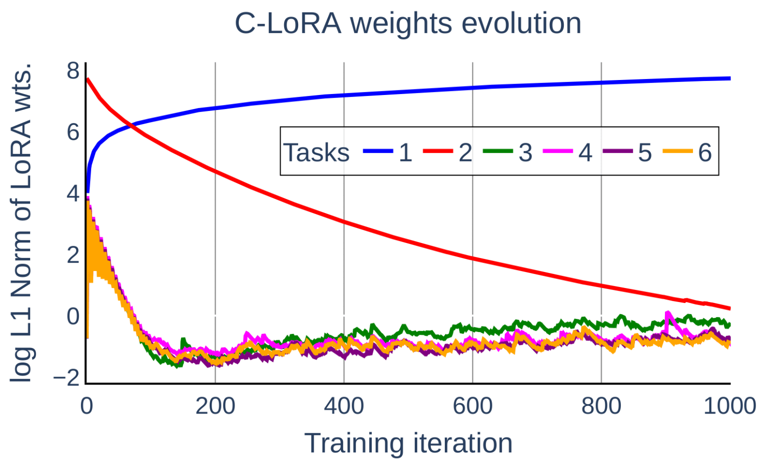
Functional-space reg. with DC scores

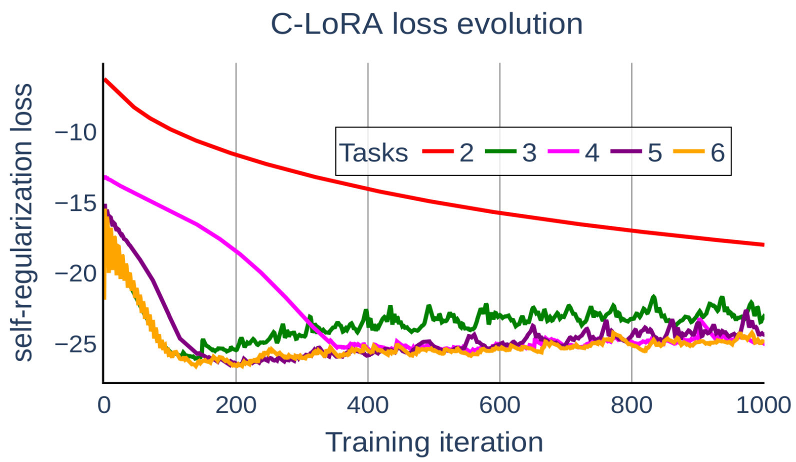
C-LoRA: SOTA parameter-space regularization

- Penalize the modification of LoRA spots allocated to any previous task

$$L_{\text{forget}} = \left\| \left| \sum_{n'=1}^{n-1} A_{n'} B_{n'} \right| \odot A_n B_n \right\|^2$$

- Leads to a degenerate solution





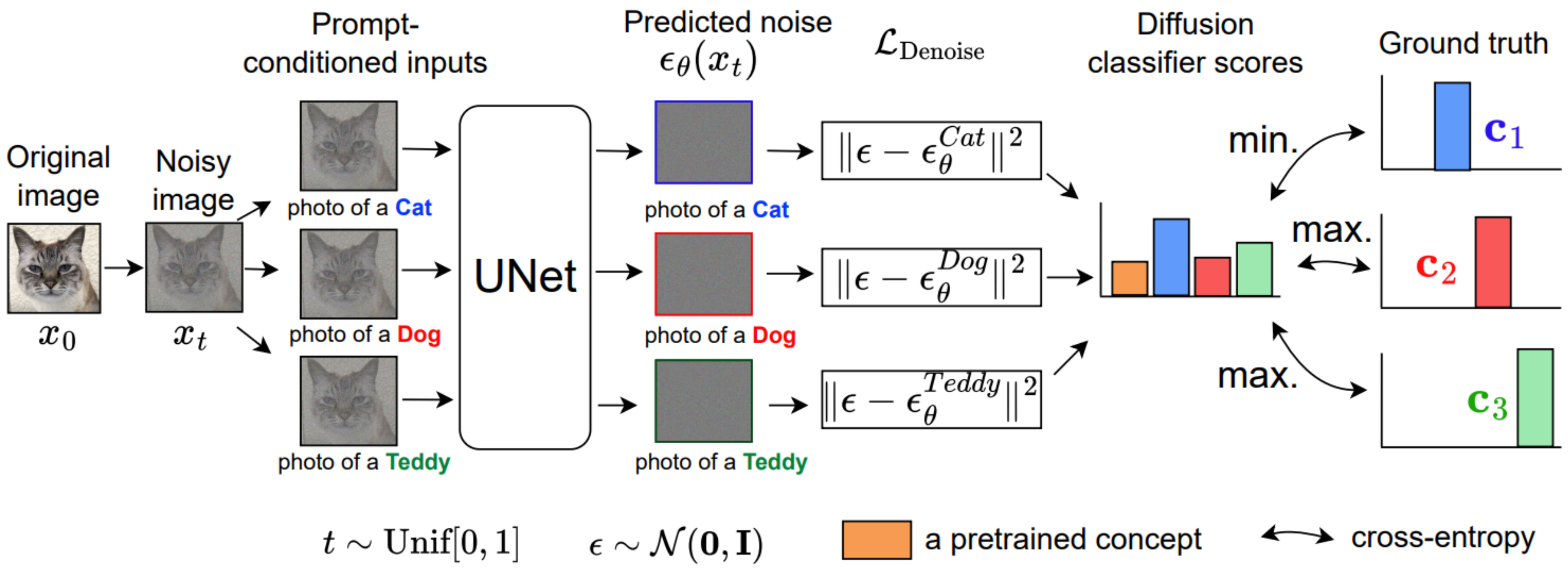
Elastic Weight Consolidation: a classic

- Parameters important to a previous task will have larger corresponding Hessian (wrt log likelihood of current task samples)
- θ^b is the model learned from previous tasks
- b_i is the diagonal Fisher information:

$$L'(\theta) = L(\theta) + \lambda \sum_i b_i(\theta_i - {\theta^b}_i)^2$$

DC scores for Fisher information estimation

- We compute the DC scores as usual
- Loss *L* for FIM estimation involves:
  - *L_Denoise* + cross-entropy(DC scores, one-hot ground truth)



Practical challenges in deriving DC scores

1. Large number of inference trials

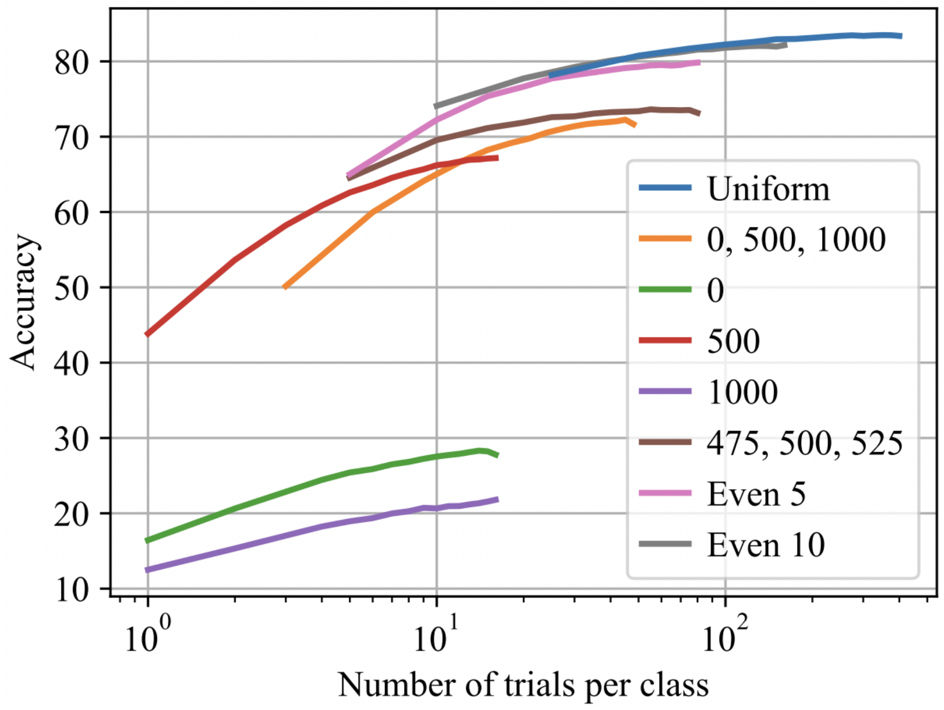


Image source: Li et al. “Your diffusion model is secretly a zero-shot classifier”, ICCV 2023.

Practical challenges in deriving DC scores

1. Large number of inference trials

- We estimate the FIM as an average over multiple epochs
- We maintain single trial per class per minibatch
- Results in diverse range of timesteps over multiple epochs

Practical challenges in deriving DC scores

1. Large number of inference trials
2. Large number of seen concepts

Practical challenges in deriving DC scores

1. Large number of inference trials
2. Large number of seen concepts
   - Iterative pruning still requires multiple passes per class
   - Instead, we maintain a subset c_k of seen concepts
   - c_k always contains c_0 (pretrained), c_n (most recent)
   - Additionally, c_k might contain $|k-2|$ randomly sampled previous concepts

Roadmap progress

EWC in LoRA space ✓

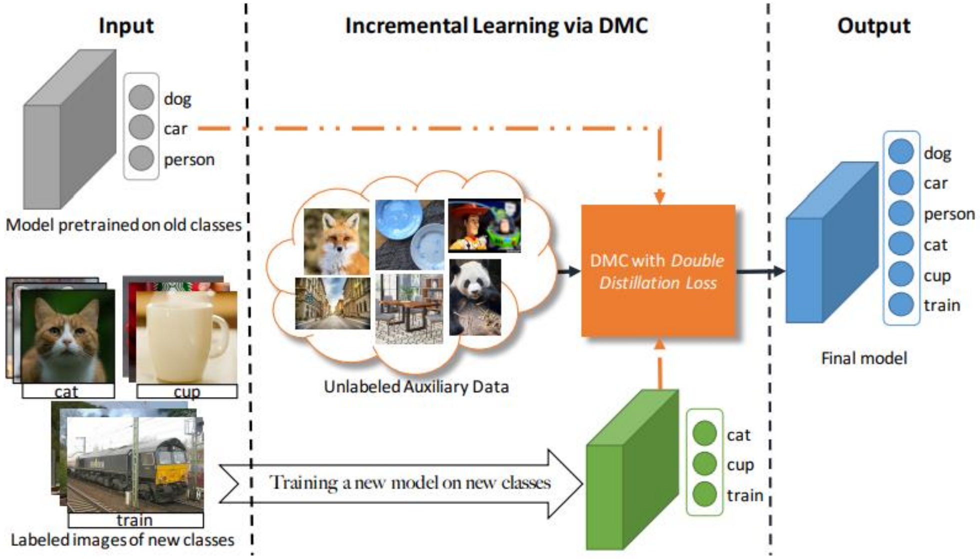
EWC with DC scores ✓

Functional-space regularization framework

Functional-space reg. with DC scores

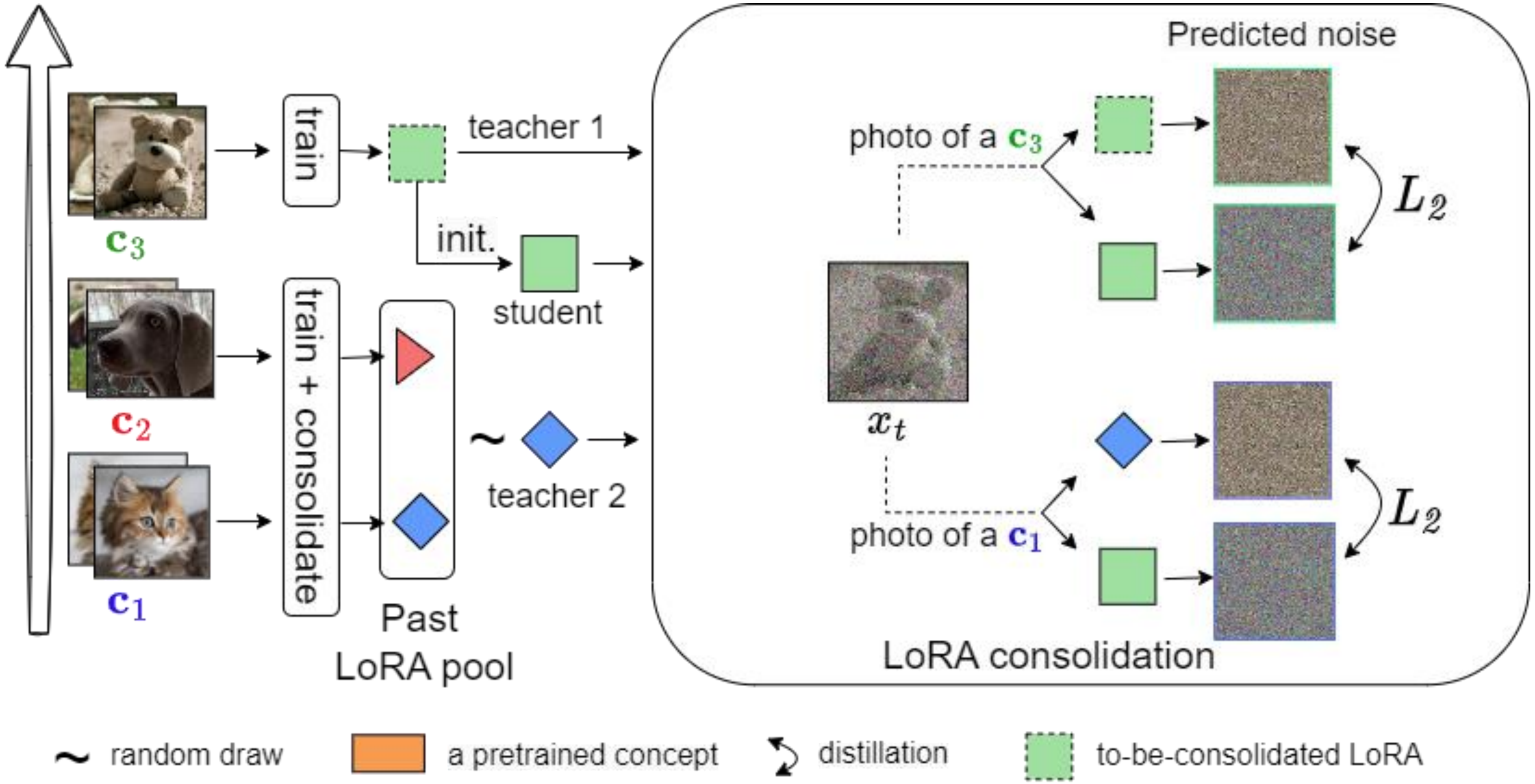
Deep Model Consolidation: another classic

- Double distillation loss for function-space consolidation
- Student model initialized randomly
- Teacher-1 model initialized from task *n*
- Teacher-2 model initialized from task *n-1*
- No need for memory replay while training student model



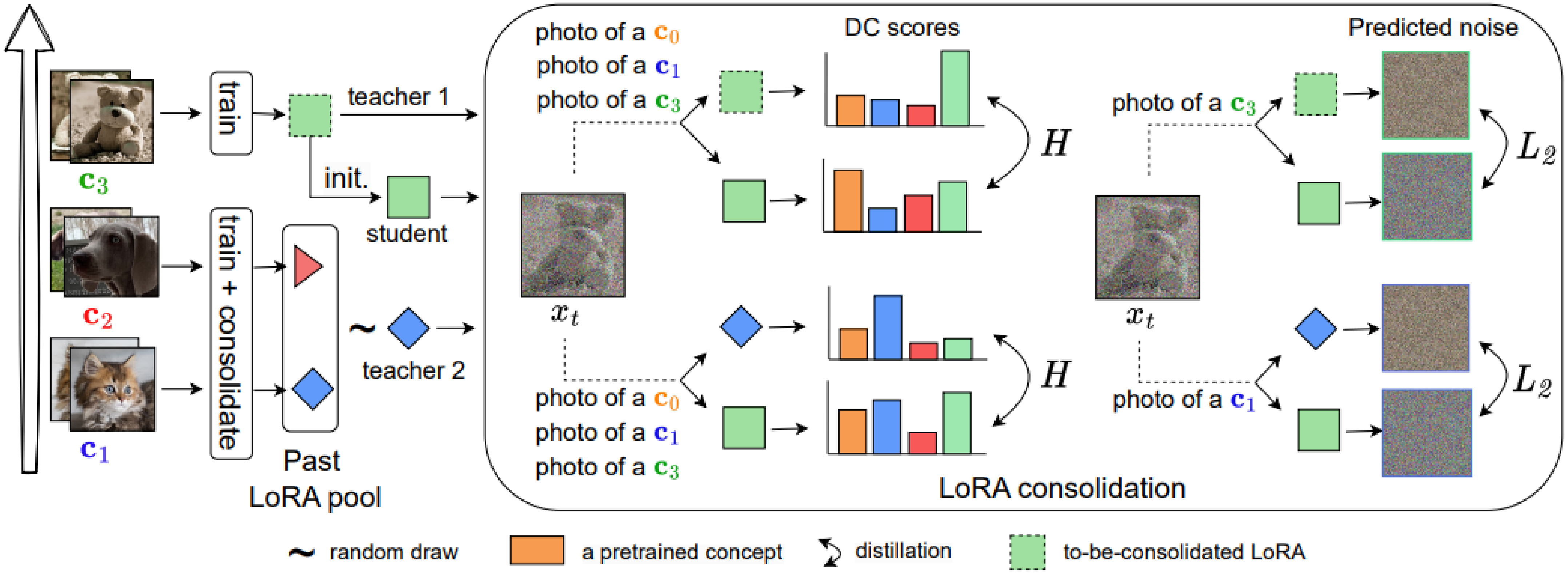
DMC for LoRA needs changes

- Teacher-1 = LoRA for task *n*
- Teacher-2 = LoRA chosen at random from tasks {1, ..., *n-1*}
- Student initialized from Teacher-1



Diffusion Scores Consolidation (DSC)

- Training objective:
  - Minimize cross-entropy between teacher-student DC scores
  - Minimize L2 distance between teacher-student noise predictions

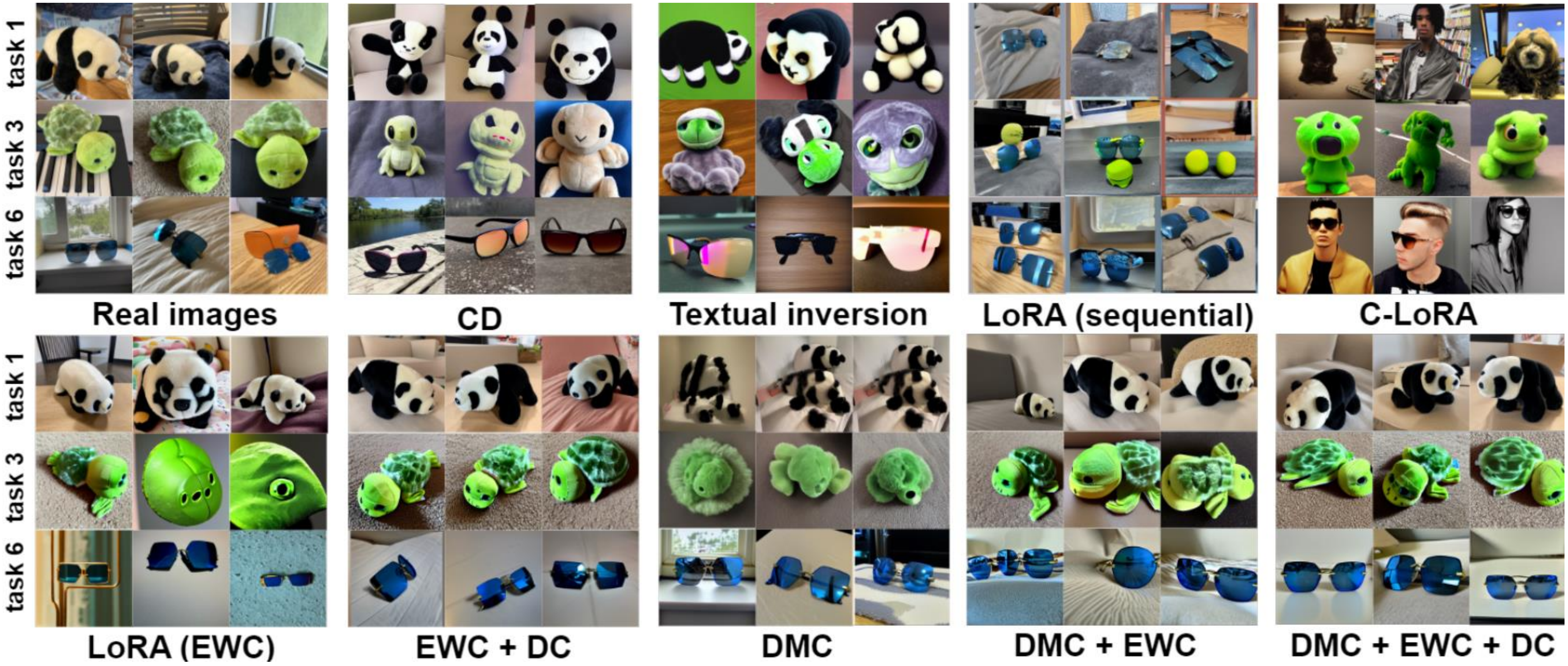


Evaluation metrics

- Using CLIP-based features for real and generated images
- The lower the better:
  - Average Maximum Mean Discrepancy (A_MMD)
  - Forgotten Maximum Mean Discrepancy (F_MMD)
  - Kernel Inception Distance (KID)
- The higher the better:
  - Image to Image similarity (I2I)
  - Text to Image similarity (T2I)
  - Backward transfer of MMD scores (BwT_MMD): Proposed to address the relative natures of F_MMD.

BWT_MMD = $\frac{1}{(N-1)} \sum_{j=1}^{N-1} (\mathrm{MMD}(F_{CLIP}(XD_{,j}), F_{CLIP}(Xj_{,j})) - \mathrm{MMD}(F_{CLIP}(X_{D,j}), FC_{LIP}(XN_{,j})))$

Qualitative results: CustomConcept dataset



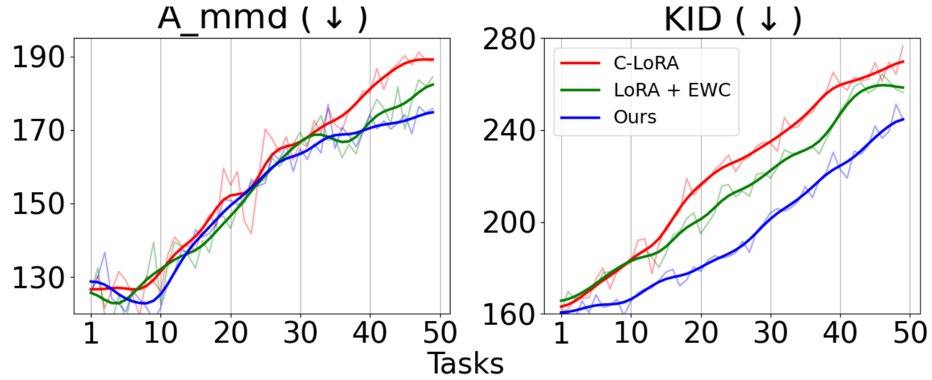
More results: https://srvcodes.github.io/continual_personalization/

Quantitative results: CustomConcept dataset

After 6 tasks (avg. over 3 seeds)	CLIP I2I (x100) ↑	CLIP T2I (x100) ↑	KID (x 10^3) ↓	A_MMD (x 10^3) ↓	Forgetting MMD ↓	BwT MMD ↑
Textual Inversion	60.74	22.86	205.69	185.74	0	0
Custom Diffusion (CD)	69.53	22.55	179.4	121.89	0.62	-273.41
CD + EWC	69.44	22.58	177.99	121.02	0.506	-245.7
CD with LoRA	61.30	22.97	203.11	176.38	0.052	-118.56
C-LoRA	64.89	**23.07**	173.8	117.2	0.034	-107.47
LoRA + EWC	73.19	22.15	156.91	105.07	0.008	-99.34
LoRA + EWC + DC	**73.41**	22.97	154.25	102.81	**0.00052**	-102.53
LoRA + DMC	73.36	22.57	187.2	198.45	0.049	-105.79
LoRA + DMC + EWC	72.92	22.89	143.92	98.0	0.02	-94.63
LoRA + DMC + EWC + DC	73.17	22.84	**140.18**	**94.1**	0.003	**-92.44**

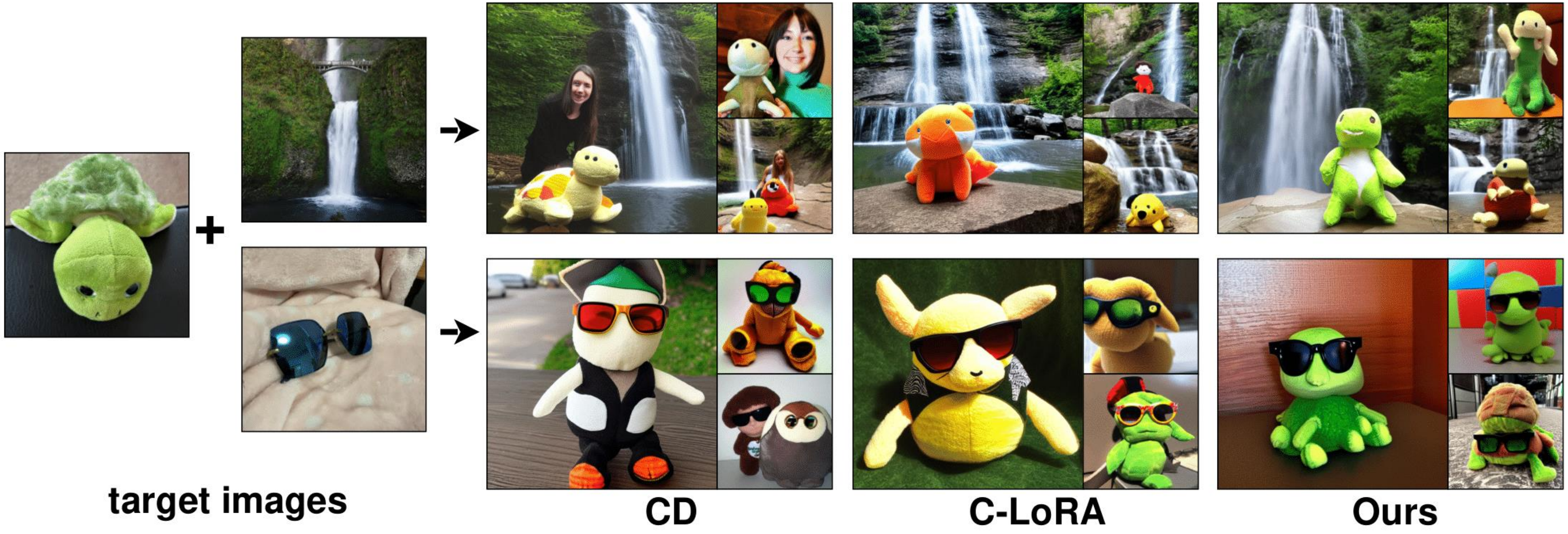
Longer sequence: 50 tasks setup

- Randomly chosen subset of 50 concepts from the CustomConcept101 dataset



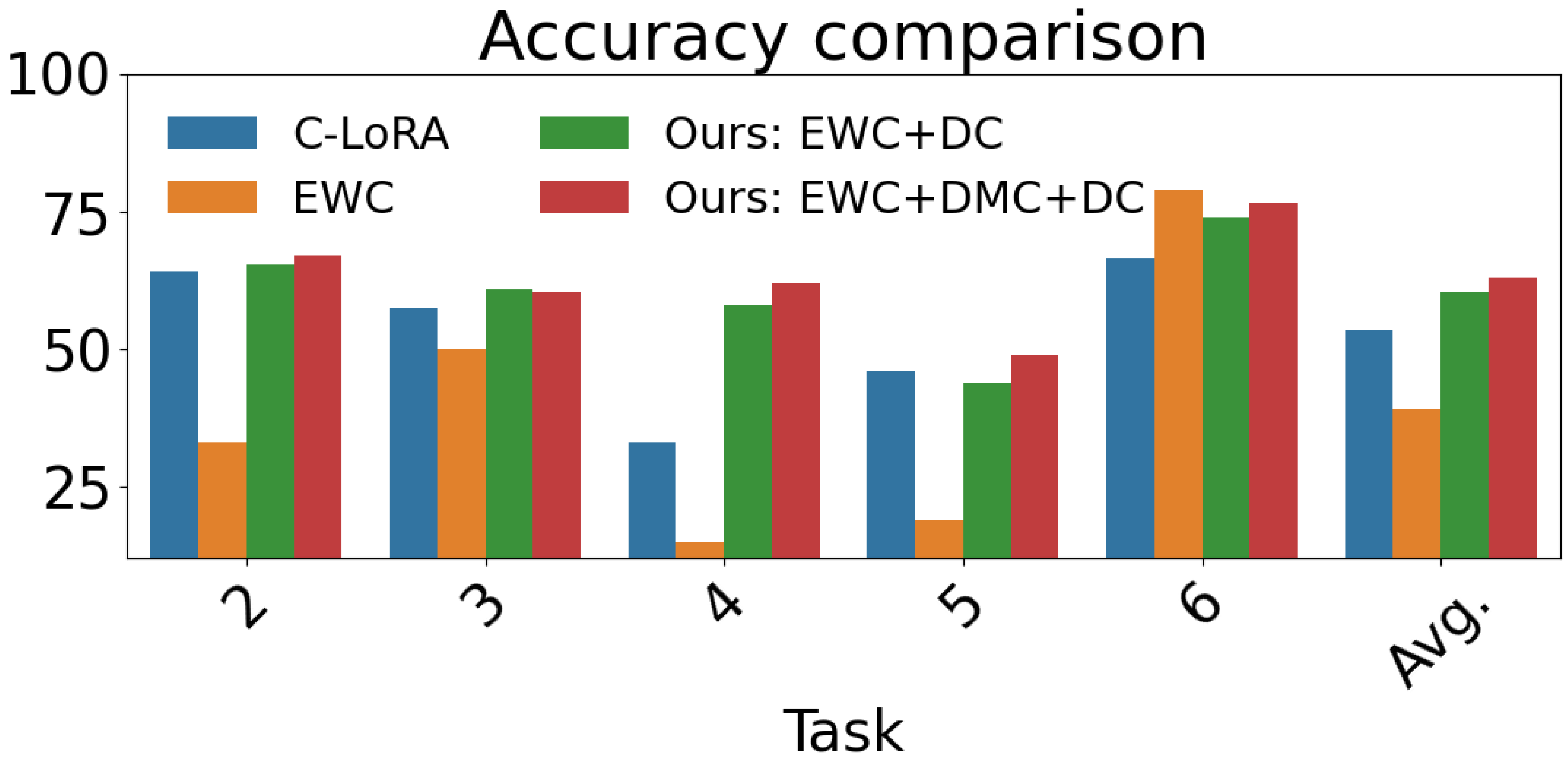
Multi-concept generation setup

- Prompt: “A photo of V1 plushie tortoise. Posing in front of V2 waterfall”



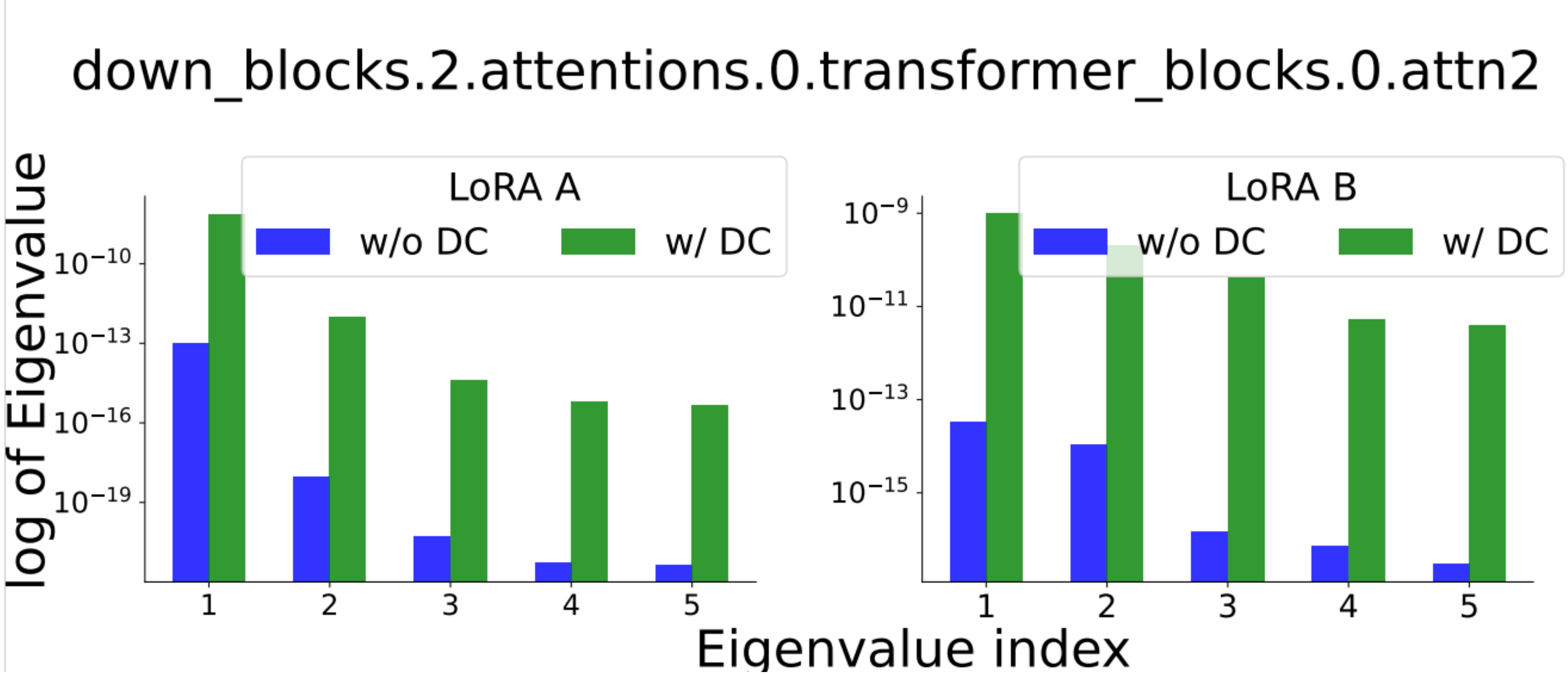
Sanity check-1 for DC scores

- Evaluating classification accuracy on the training dataset



Sanity check-2 for DC scores

- Evaluating the information encoded in the Fisher Information Matrix
- Choose 3 random layers, and check the top-k Eigen values for FIM



Conclusion

- We study using DC scores for continual personalization of text-to-image diffusion model
- We propose two regularization frameworks for DC scores:
  - Parameter-space reg. with Elastic Weight Consolidation
  - Function-space reg. with Deep Model Consolidation
- Both proposed methods have zero inference-time parameter overhead over state-of-the-art C-LoRA