How to Continually Adapt Text-to-Image Diffusion Models for Flexible Customization? (NeurIPS 2024)

Jiahua Dong[1*], Wenqi Liang[2*], Hongliu Li[3], Duzhen Zhang[1], Meng Cao[1], Henghui Ding[4], Salman Khan[1,5], Fahad Shahbaz Khan[1,6]

[1]Mohamed bin Zayed University of Artificial Intelligence

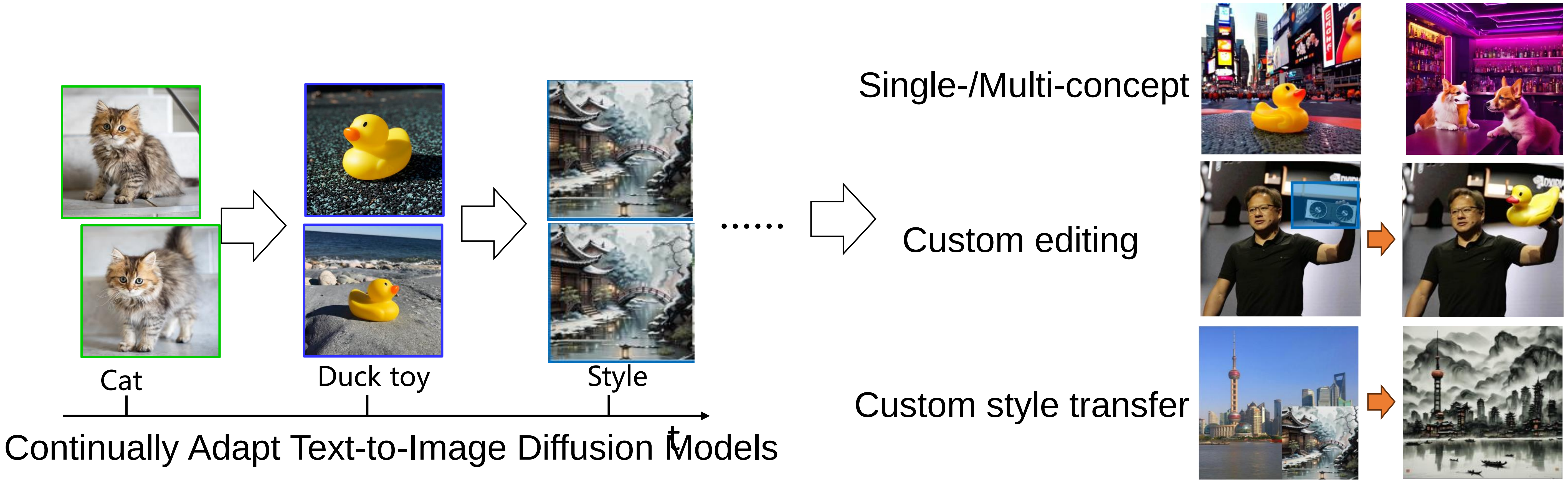
[2]Shenyang Institute of Automation, Chinese Academy of Sciences,

[3]The Hong Kong Polytechnic University, [4]Institute of Big Data, Fudan University

[5]Australian National University, [6]Linköping University

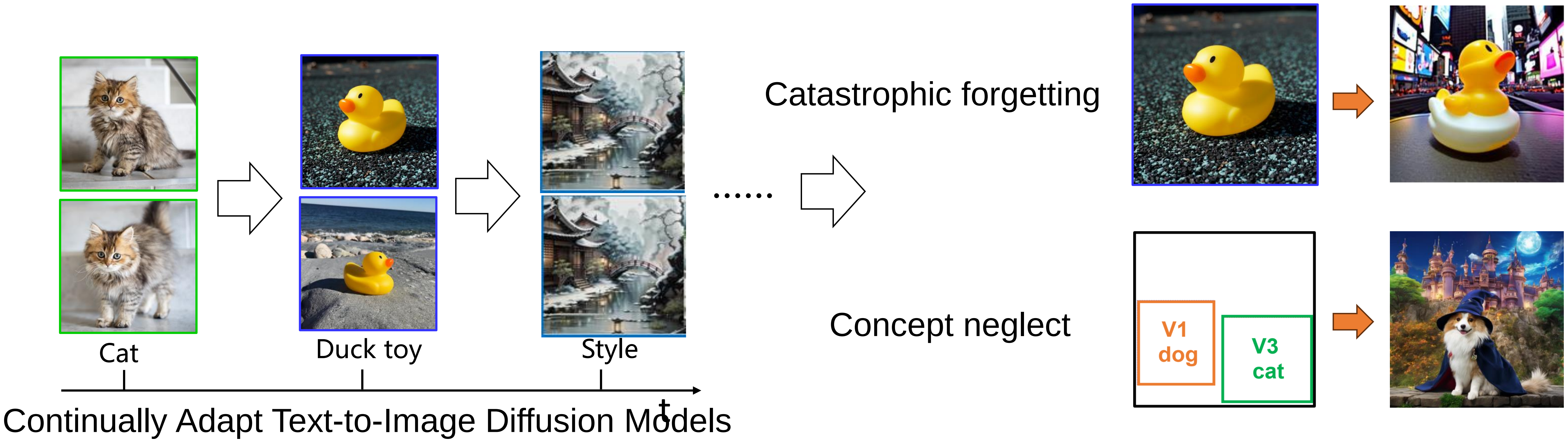
◆ Background

- Concept-Incremental Diffusion: Continually synthesize a series of new personalized concepts from user's own lives (i.e., pets, objects, style photos and human photos).
- Versatile concept customization: Consecutively synthesize a sequence of new personalized concepts for versatile customization (e.g., multi-concept generation, style transfer and image editing).



◆ Motivation

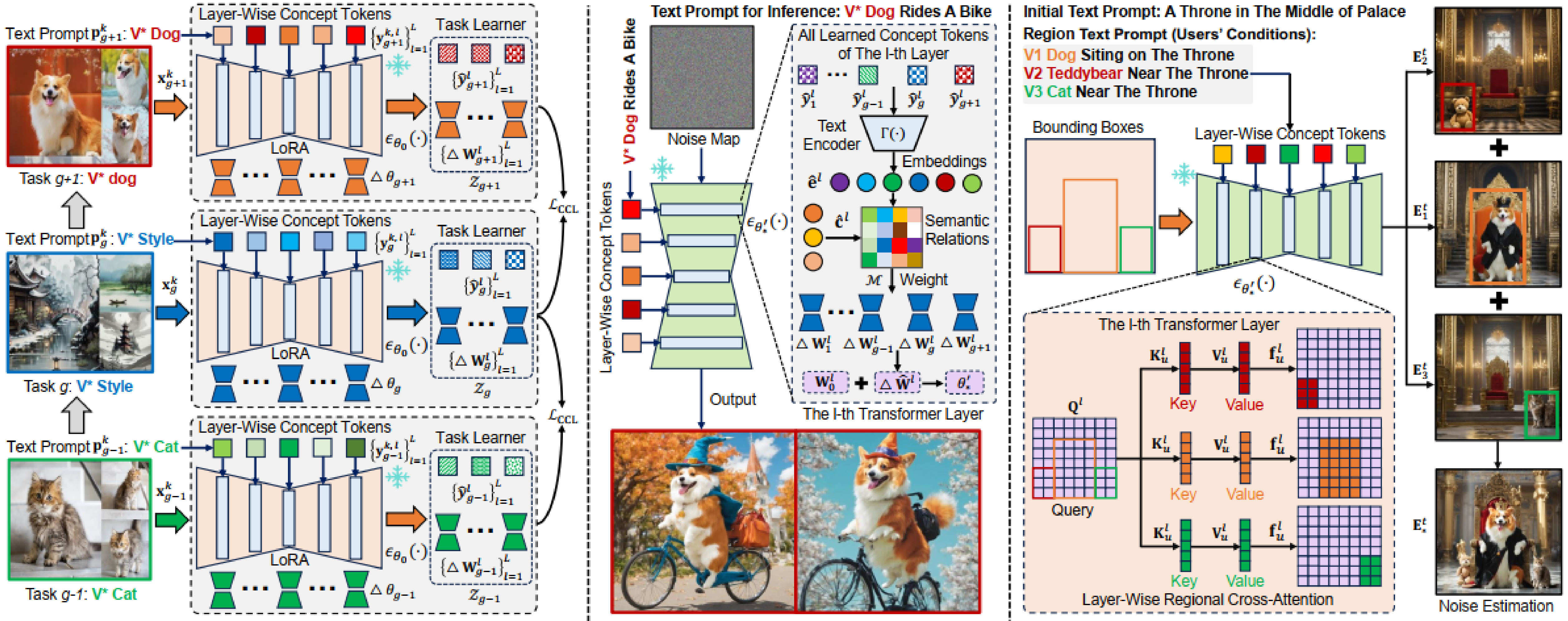
- Retain all lora weights associated with old concepts and then merge them, which may experience significant loss of individual attributes (i.e., catastrophic forgetting) for versatile customization.
- Current methods heavily suffer from concept neglect when users may wish to control the contexts and objects associated with multiple concepts in synthesized images.



How to Continually Adapt Text-to-Image Diffusion Models for Flexible Customization?

Contributions:

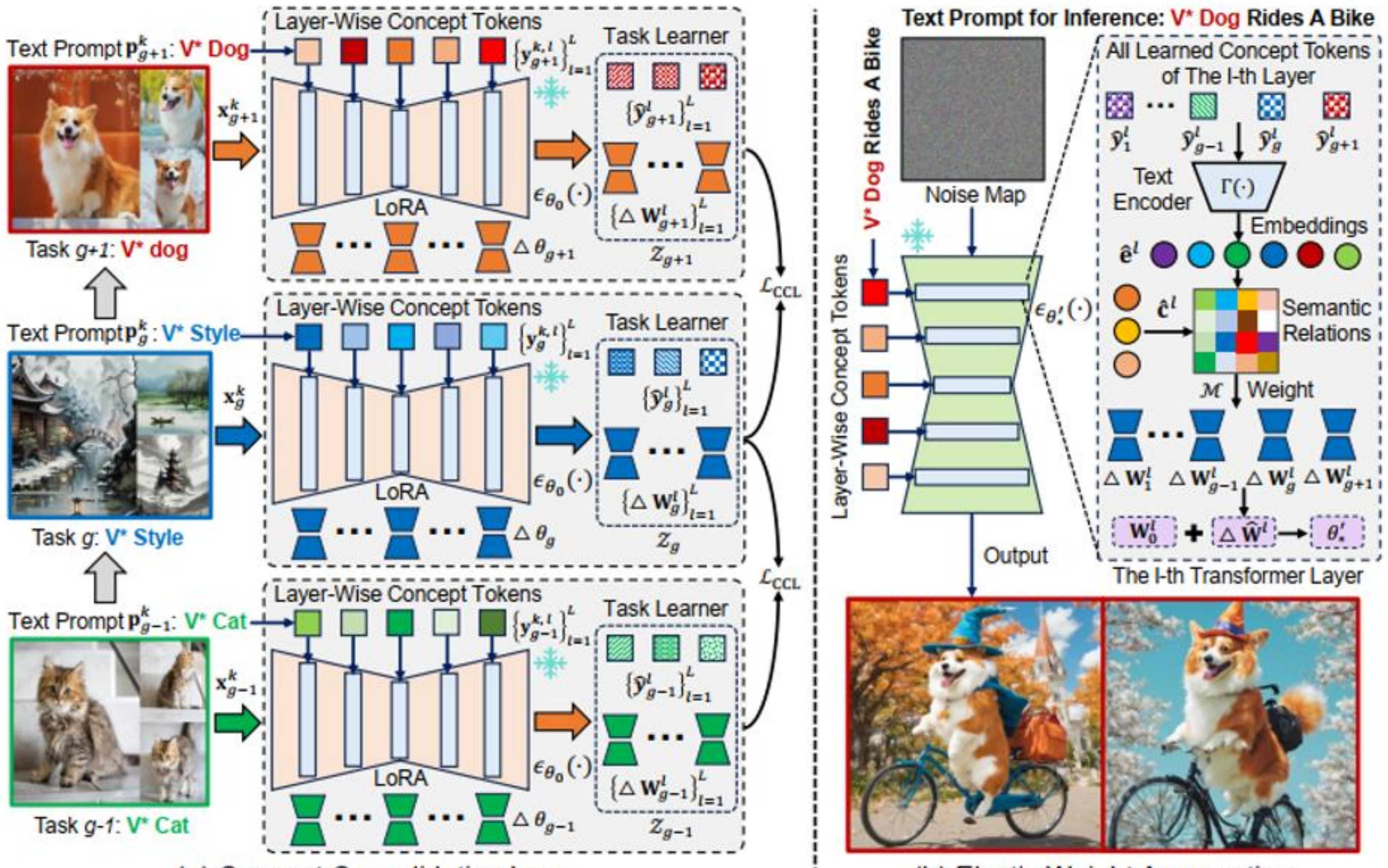
- Develop a novel Concept-Incremental text-to-image Diffusion Model (CIDM) to learn new personalized concepts continuously for versatile concept customization.
- Devise a concept consolidation loss and an elastic weight aggregation module to mitigate the catastrophic forgetting.
- Develop a context-controllable synthesis strategy to tackle the concept neglect.



(a) Concept Consolidation Loss (b) Elastic Weight Aggregation (c) Context-Controllable Synthesis

How to Continually Adapt Text-to-Image Diffusion Models for Flexible Customization?

Tackle catastrophic forgetting: Concept consolidation loss and elastic weight aggregation



(a) Concept Consolidation Loss
(b) Elastic Weight Aggregation

1. In the g-th task, we devise an orthogonal subspace regularizer to constrain the low-rank weights of different customization tasks:

$$\triangle\theta_g = \{\triangle \mathbf{W}_g^l\}_{l=1}^L \qquad \boxed{\triangle \mathbf{W}_g^l} = \mathbf{A}_g^l \mathbf{B}_g^l$$

LoRA weights of l-th layers

We perform the orthogonal subspace regularizer on the low-rank concept subspaces of different tasks:

$$\sum_{i=1}^{g-1}\sum_{l=1}^{L} \boxed{\mathbf{A}_i^l(\mathbf{A}_g^l)^\top} = 0. \; \mathcal{R}_1 = \sum_{i=1}^{g-1}\sum_{l=1}^{L} \boxed{\mathbf{A}_i^l(\mathbf{A}_g^l)^\top}$$

2. After leraning g tasks, we develop an elastic weight aggregation(EWA) module to adaptively merge them for versatile concept customization:

$$\mathcal{M} = \max(\boxed{\hat{\mathbf{c}}^l} \cdot \boxed{(\hat{\mathbf{e}}^l)}^\top), \quad \triangle\widehat{\mathbf{W}}^l = \sum_{i=1}^{g} \boxed{\triangle \mathbf{W}_i^l}\, \psi(\mathcal{M})_i,$$

Layer-wise concept embeddings

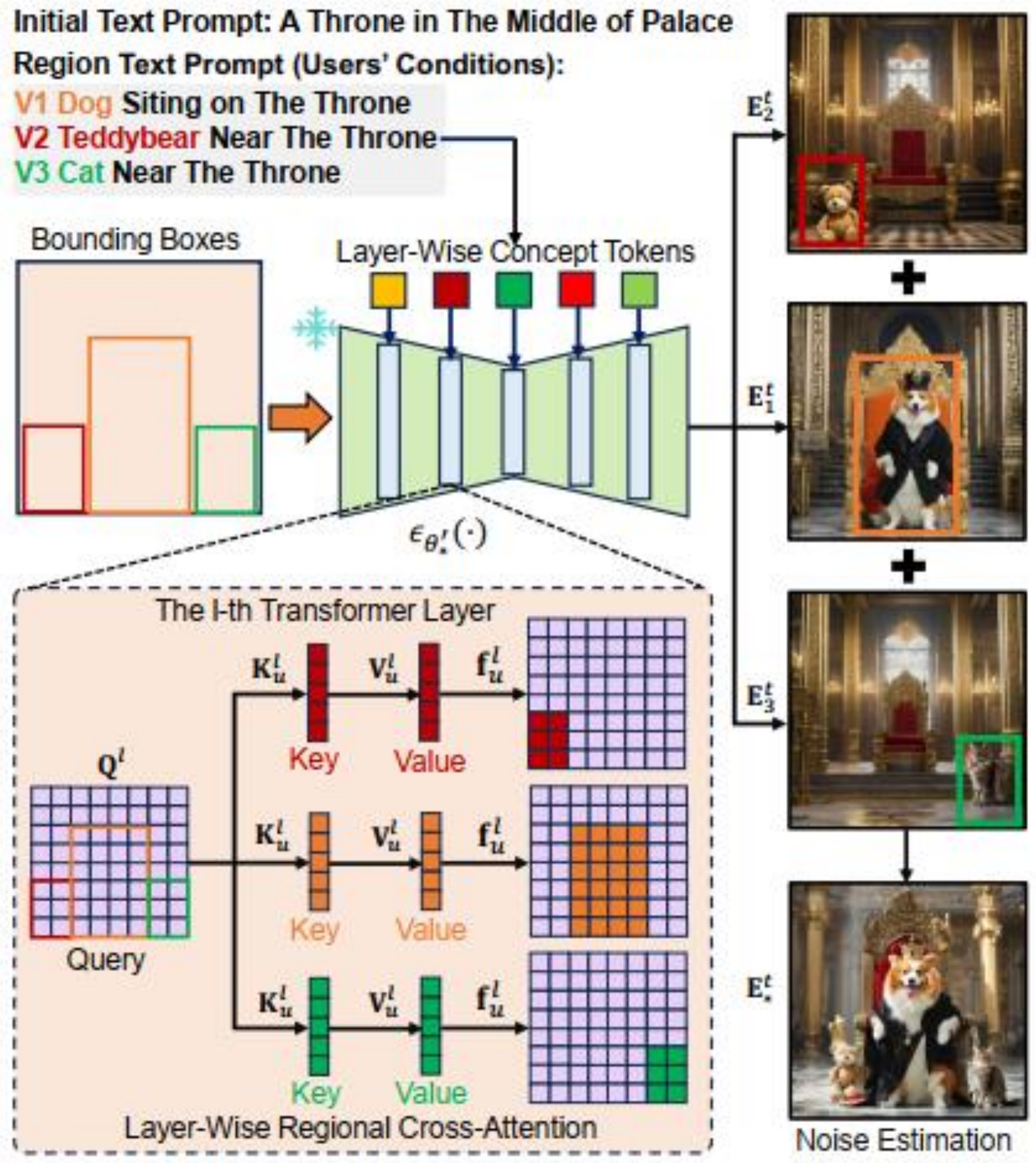
Layer-wise text embeddings

All learned lora weights

How to Continually Adapt Text-to-Image Diffusion Models for Flexible Customization?

NEURAL INFORMATION PROCESSING SYSTEMS

Tackle concept neglect : Context-controllable synthesis strategy



(c) Context-Controllable Synthesis

Current methods suffer catastrophic neglect when generating images of multi-concepts.

1. Perform layer-wise regional cross-attention between textual embedding and latent feature for i-th region:

$$\mathbf{Q}^l = \Omega(\mathbf{f}^l \mathbf{w}_q \odot \boxed{\widehat{\mathbf{m}}_u^l})$$

Region mask

$$\mathbf{K}_u^l = \widehat{\mathbf{c}}_u^l \mathbf{w}_k \in \mathbb{R}^{n_e \times d}$$

$$\mathbf{V}_u^l = \widehat{\mathbf{c}}_u^l \mathbf{w}_v \in \mathbb{R}^{n_e \times d}$$

2. we aggregate U regional noise estimations to further address concept neglect.

$$\boxed{\mathbf{E}_u^t} = \epsilon_{\theta'_*}(\mathbf{z}_t|t) + s \cdot (\epsilon_{\theta'_*}(\mathbf{z}_t|[\widehat{\mathbf{c}}_u, \widehat{\mathbf{s}}_u], t) - \epsilon_{\theta'_*}(\mathbf{z}_t|t)),$$

Forward noise estimations

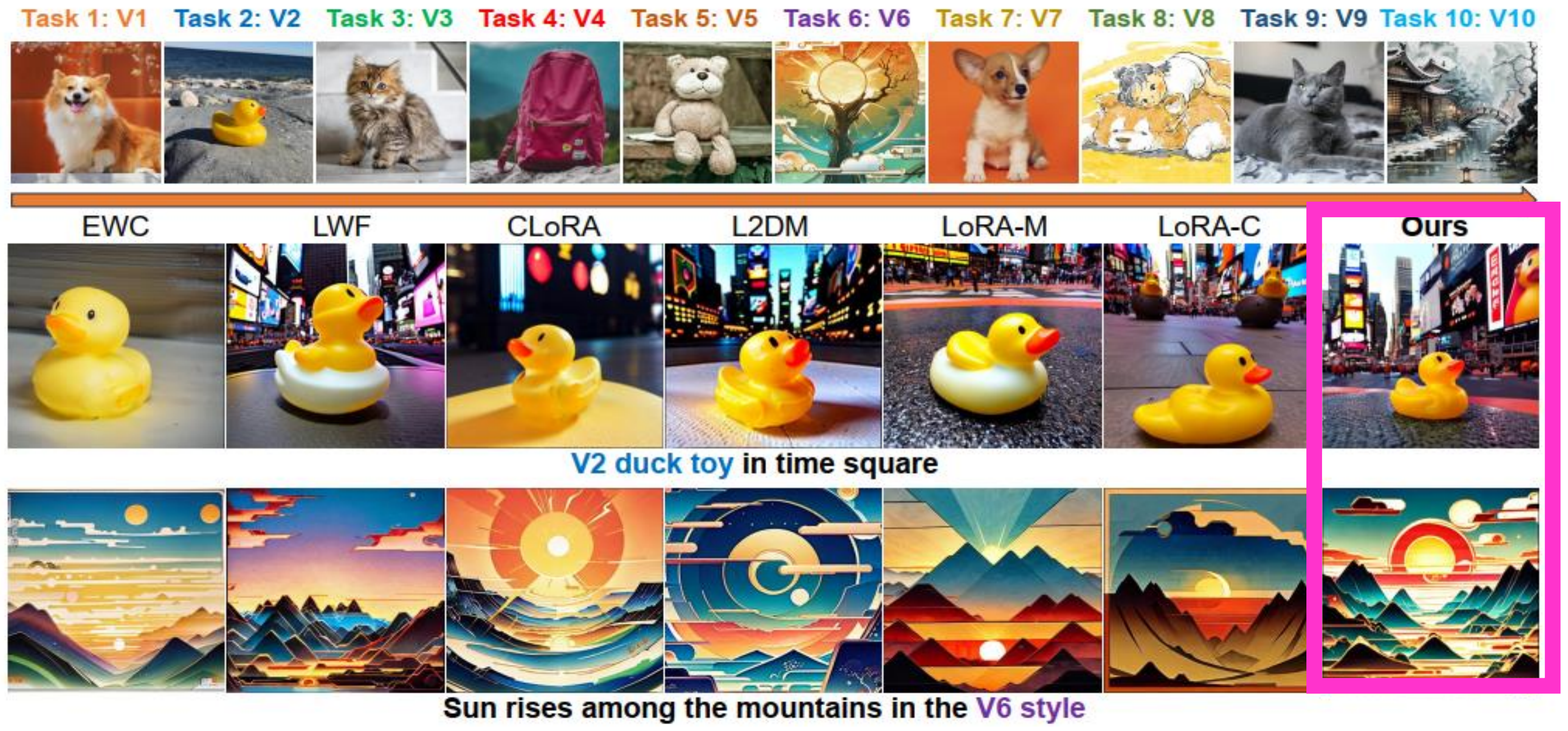
$$\boxed{\mathbf{E}_*^t} = \alpha \mathbf{E}^t + \sum_{u=1}^{U} (1-\alpha) \boxed{\mathbf{E}_u^t} \odot \widehat{\mathbf{m}}_u^L,$$

How to Continually Adapt Text-to-Image Diffusion Models for Flexible Customization?

◆ Experiments： Concept-incremental settings

Datasets:

Single-concept:



Multi-concept:

How to Continually Adapt Text-to-Image Diffusion Models for Flexible Customization?

◆ Experiments： Concept-incremental settings

Qualitative Comparisons:

Achieve **2.0%~8.0%** improvement

Table 1: Comparisons (IA) of single-concept customization synthesized by SD-1.5 and SDXL.

Methods	SD-1.5 [40]											SDXL [35]										
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	Avg.	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	Avg.
Finetuning	77.6	82.2	79.0	77.6	79.6	62.9	71.5	53.7	81.4	72.1	73.7	62.0	70.8	79.1	73.4	76.4	67.5	76.8	57.4	77.1	74.8	71.5
EWC [20]	78.7	83.8	80.4	80.3	80.7	64.0	**76.5**	57.1	**84.4**	73.1	75.9	83.6	80.5	84.6	80.8	79.2	70.1	80.5	61.2	**79.5**	75.8	77.6
LWF [26]	80.4	79.7	80.9	77.4	80.9	61.8	73.2	53.5	78.1	74.7	74.1	84.0	81.2	84.2	81.7	79.7	68.1	77.1	60.1	76.3	72.7	76.5
LoRA-M [70]	80.0	84.2	79.1	76.5	82.7	65.7	70.1	54.7	79.5	74.1	74.6	82.6	79.9	84.5	80.1	80.9	57.8	77.0	54.0	71.8	74.0	74.3
LoRA-C [70]	80.1	84.1	79.8	76.6	82.9	65.9	70.8	54.9	79.9	74.4	74.9	82.8	80.4	84.8	80.0	81.0	58.2	76.8	54.5	72.2	73.9	74.5
CLoRA [46]	83.2	83.4	81.1	80.6	84.9	66.3	76.2	58.1	83.0	72.1	76.9	83.4	81.3	85.8	80.1	79.0	**70.4**	81.2	61.7	78.5	**76.7**	77.8
L2DM [48]	78.7	86.3	76.6	80.7	**86.8**	**70.8**	70.0	**59.3**	77.7	74.1	76.1	84.6	79.5	81.9	75.5	82.1	69.2	80.9	**63.8**	77.0	76.4	77.1
CIDM (Ours)	**83.6**	**86.4**	**82.9**	**80.8**	86.5	69.5	73.7	56.9	82.4	**75.9**	**78.0**	**87.1**	**82.1**	**88.5**	**84.9**	**85.8**	68.3	**82.0**	62.4	76.9	76.6	**79.5**

Table 2: Comparisons (TA) of single-concept customization synthesized by SD-1.5 and SDXL.

Methods	SD-1.5 [40]											SDXL [35]										
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	Avg.	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	Avg.
Finetuning	64.4	74.6	69.4	68.6	75.0	70.0	**76.7**	69.2	65.4	67.2	70.0	54.8	77.5	72.2	85.0	80.5	76.2	**79.7**	73.6	77.6	76.3	75.3
EWC [20]	67.1	77.5	72.7	77.9	76.7	**72.3**	74.2	72.0	66.0	70.4	72.7	71.4	79.8	72.8	84.4	79.5	73.9	76.7	77.0	78.3	77.6	77.1
LWF [26]	70.8	75.2	71.0	77.4	76.0	71.7	76.3	72.9	72.5	70.0	73.4	**75.8**	76.9	76.0	83.6	82.9	75.1	76.7	74.3	79.1	76.8	77.7
RPY [27]	68.1	76.2	70.1	78.4	75.7	69.3	74.8	70.5	65.8	68.6	71.8	69.3	**81.0**	71.9	87.3	78.8	71.5	76.4	75.9	79.7	76.2	76.8
CLoRA [46]	69.4	78.0	**74.1**	78.8	76.4	69.6	**76.7**	73.9	69.0	**71.8**	73.6	71.8	80.1	71.1	**87.7**	81.2	74.6	77.8	77.7	80.1	75.9	77.8
L2DM [48]	68.6	**79.5**	70.1	73.0	76.7	67.7	75.9	74.1	71.8	69.4	72.7	72.6	78.4	**78.5**	85.0	81.5	73.5	78.6	79.1	**81.9**	77.8	78.7
CIDM (Ours)	**75.3**	78.1	74.0	**81.1**	**78.2**	70.1	74.7	**74.3**	**73.5**	70.2	**74.8**	74.9	79.6	74.5	86.7	**83.5**	**79.8**	78.2	**83.1**	81.4	**78.5**	**80.0**

Thanks for your attention!

Code Link: https://github.com/JiahuaDong/CIFC

Email: dongjiahua1995@gmail.com

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

香港城市大學 City University of Hong Kong

LINKÖPING UNIVERSITY