中国人民大學
RENMIN UNIVERSITY OF CHINA

高瓴人工智能学院
Gaoling School of Artificial Intelligence

Bridging The Gap between Low-rank and Orthogonal Adaptation via Householder Reflection Adaptation

Shen Yuan, Haotian Liu, Hongteng Xu

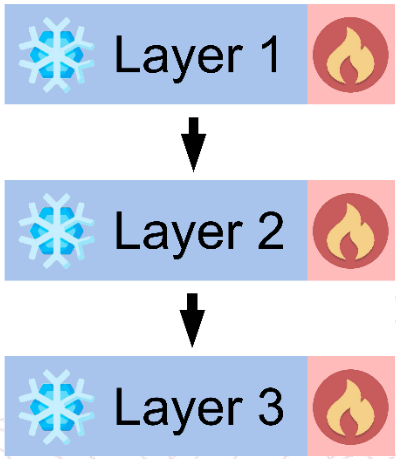
Gaoling School of Artificial Intelligence, Renmin University of China

{shenyuan721, haotianliu, hongtengxu}@ruc.edu.cn

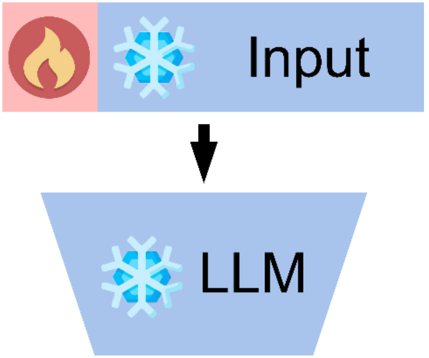
NeurIPS 2024 [Spotlight]

Parameter-efficient fine-tuning (PEFT)

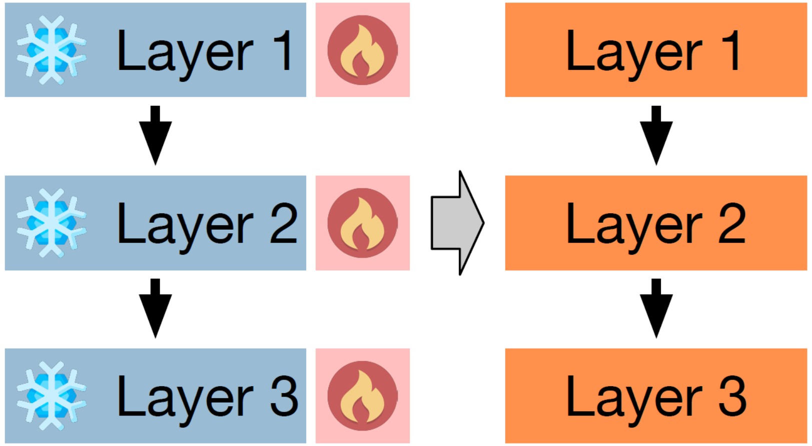
- **Model fine-tuning**



- **Soft prompt fine-tuning**

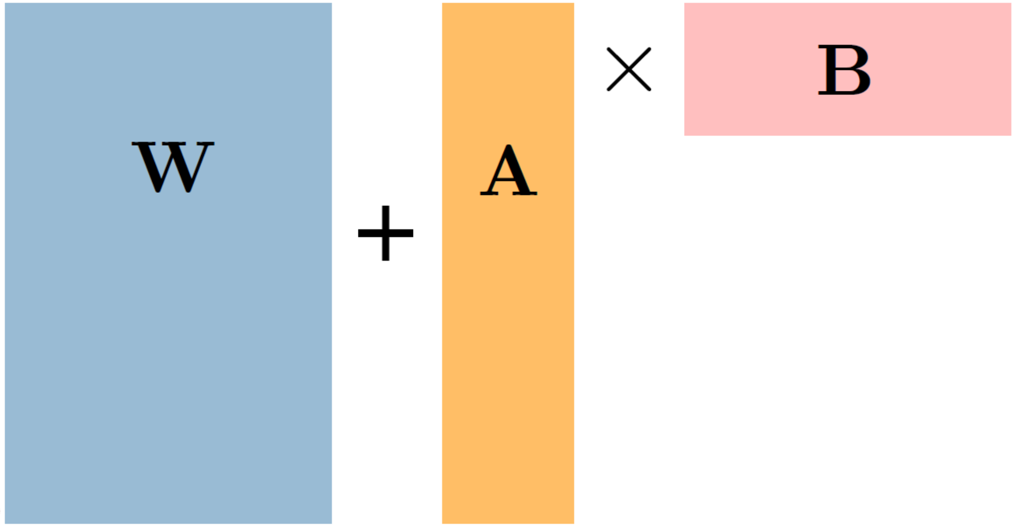


- **Adapter-based fine-tuning**



LoRA: The mainstream model adaptation method

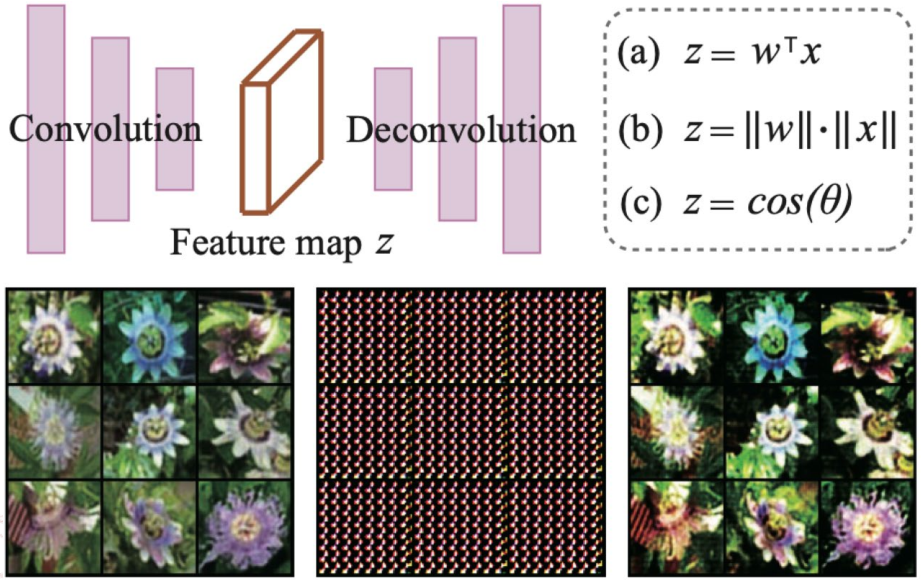
- LoRA hypothesizes that the weights have a low “intrinsic rank” during adaptation.



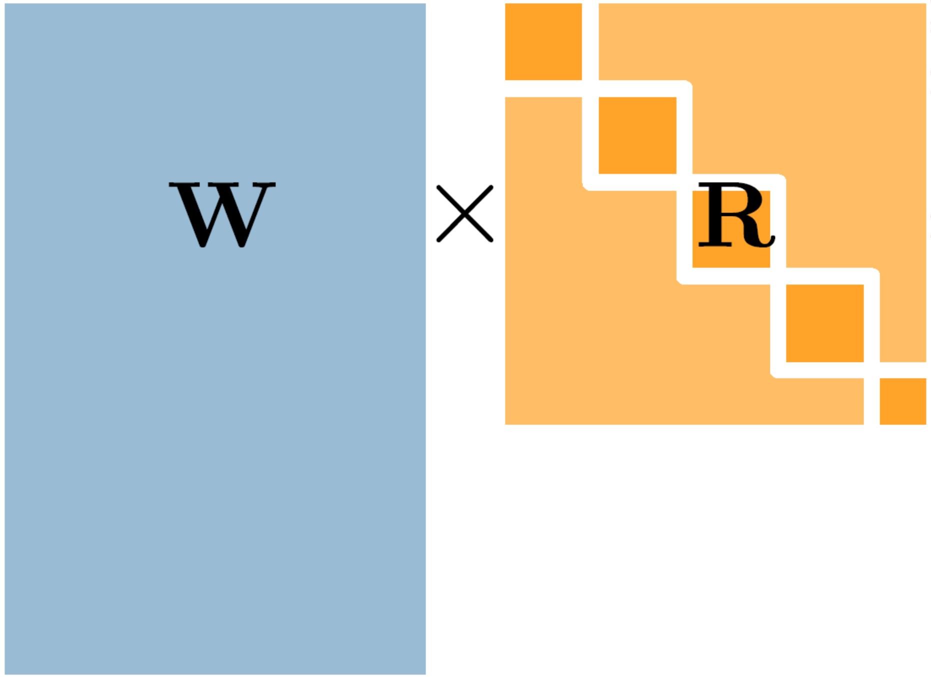
Hu, Edward J., et al. ”LoRA: Low-Rank Adaptation of Large Language Models.” ICLR 2022.

OFT: The preservation of pretrained knowledge

- OFT emphasizes the retention of pre-trained knowledge during adaptation.



(a) Inner product (b) Magnitude (c) Angle

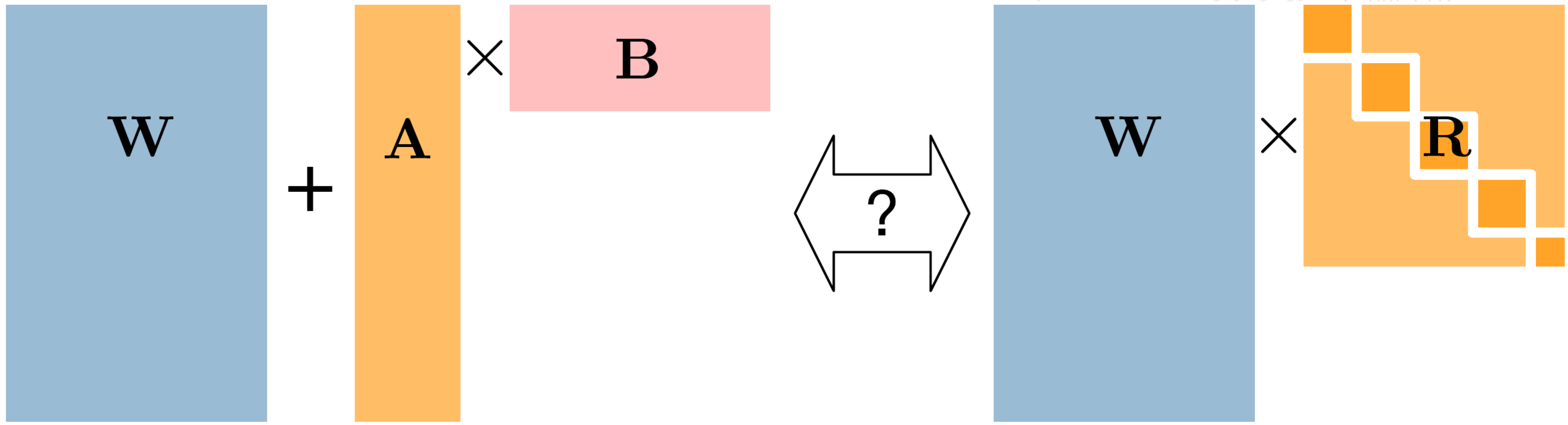


Angular information matters

→ Orthogonal fine-tuning (OFT)

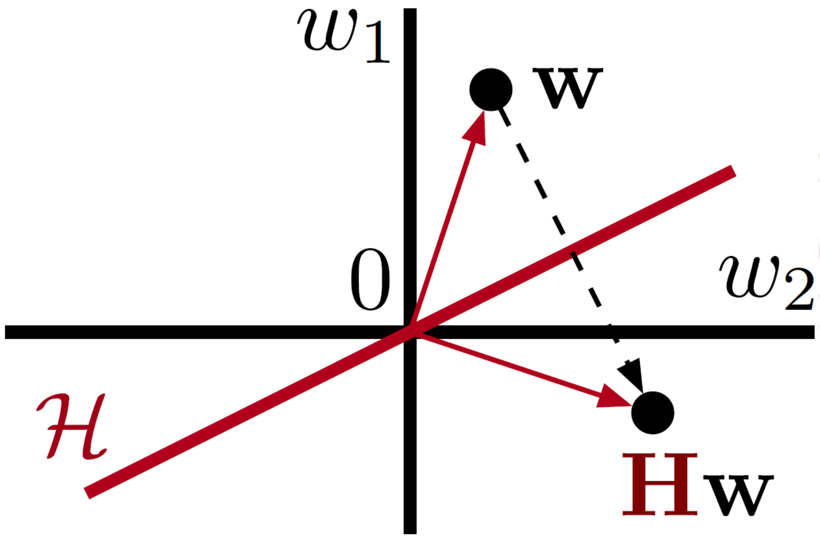
Qiu, Zeju, et al. "Controlling text-to-image diffusion by orthogonal finetuning." NeurIPS 2023.

Is there a bridge between LoRA and OFT?



Householder Reflection: A simple orthogonal transform

Alston Householder

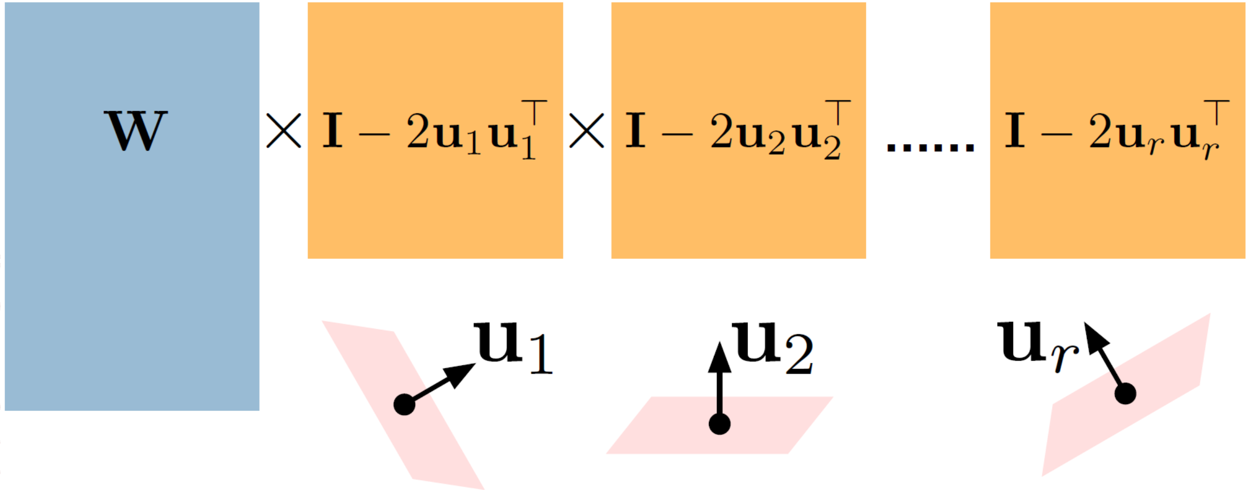


$$\boldsymbol{H} = \boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^\top, \quad \boldsymbol{u} \in \mathbb{S}^{d-1}$$

Householder Reflection Adaptation (HRA)

- Implement OFT by a chain of Householder reflections

$$\boldsymbol{z} = \boldsymbol{W}\underbrace{\left(\prod\nolimits_{i=1}^{r}\boldsymbol{H}_i\right)}_{\boldsymbol{H}^{(r)}}\boldsymbol{x} = \boldsymbol{W}\left(\prod\nolimits_{i=1}^{r}(\boldsymbol{I} - 2\boldsymbol{u}_i\boldsymbol{u}_i^{\top})\right)\boldsymbol{x},\ \text{with } \{\boldsymbol{u}_i \in \mathbb{S}^{d-1}\}_{i=1}^{r}.$$



- Implement $\boldsymbol{W}\boldsymbol{H}^{(r)}$ with low complexity ($\mathcal{O}(d(r + d_{\text{out}}))$ for $\boldsymbol{W} \in \mathbb{R}^{d_{\text{out}}\times d}$)

1) $\boldsymbol{x}^{(j+1)} = \boldsymbol{x}^{(j)} - 2\langle \boldsymbol{u}_{r-j}, \boldsymbol{x}^{(j)}\rangle \boldsymbol{u}_{r-j}$, for $j = 0, \ldots, r-1$. 2) $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x}^{(r)}$.

Connections to LoRA: HRA is an adaptive LoRA

- Reformulation of the HR chain:

$$\boldsymbol{H}^{(r)} = \prod_{i=1}^{r}(\boldsymbol{I} - 2\boldsymbol{u}_i\boldsymbol{u}_i^{\top}) = \boldsymbol{I} + \boldsymbol{U}_r\boldsymbol{\Gamma}_r\boldsymbol{U}_r^{\top},$$

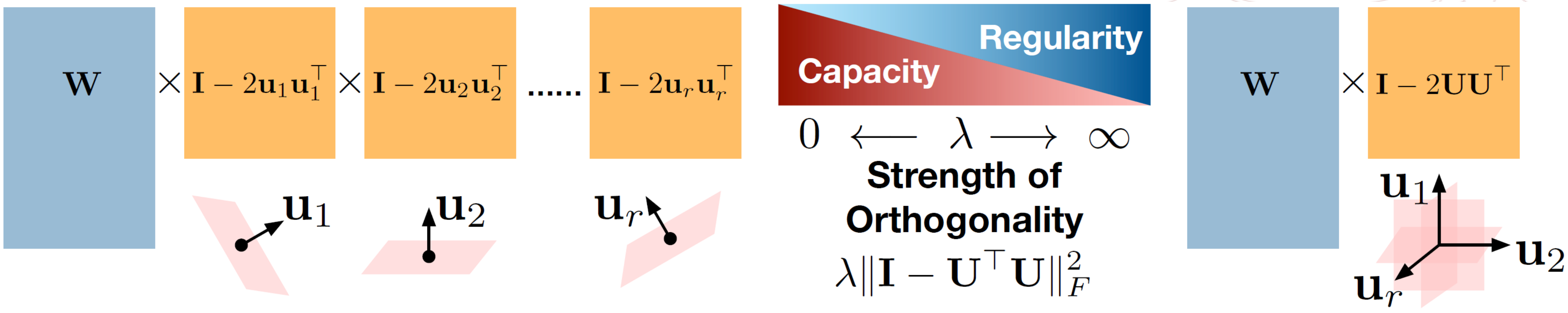
- $\boldsymbol{\Gamma}_r = [\gamma_{ij}] \in \mathbb{R}^{r\times r}$ is a upper-triangular matrix, and its upper-triangular element is

$$\boldsymbol{\Gamma}_1 = -2,\ \boldsymbol{\Gamma}_r = \begin{bmatrix} \boldsymbol{\Gamma}_{r-1} & -2\boldsymbol{\Gamma}_{r-1}\boldsymbol{U}_{r-1}^{\top}\boldsymbol{u}_r \\ \boldsymbol{0}_{r-1}^{\top} & -2 \end{bmatrix}.$$

- HRA is equivalent to an adaptive LoRA, making $Range(\boldsymbol{W})$ unchanged.

$$\boldsymbol{W}\boldsymbol{H}^{(r)} = \boldsymbol{W} + \underbrace{\boldsymbol{W}\boldsymbol{U}_r\boldsymbol{\Gamma}_r}_{\boldsymbol{A}_{\boldsymbol{W},\boldsymbol{U}}}\boldsymbol{U}_r^{\top}.$$

Orthogonality: The key of balancing expressiveness and regularity



$$\min_{\{\boldsymbol{U}_r^{(l)}\}_{l=1}^{L}} \mathrm{Loss}(\mathcal{D}; \{\boldsymbol{U}_r^{(l)}\}_{l=1}^{L}) + \underbrace{\lambda \sum_{l=1}^{L} \|\boldsymbol{I}_r - (\boldsymbol{U}_r^{(l)})^\top \boldsymbol{U}_r^{(l)}\|_F^2}_{\text{Orthogonal regularizer}},$$

- $\lambda \in [0, \infty)$: Normalization,
- $\lambda = \infty$: (Modified) Gram-Schmidt (GS) Orthogonalization.

Experiments: NLP tasks

Table: Results (%) of various methods on GLUE development set.

Method	#Param (M)	MNLI	SST-2	CoLA	QQP	QNLI	RTE	MRPC	STS-B	All
Full Fine-tune	184	89.90	95.63	69.19	**92.40**	94.03	83.75	89.46	91.60	88.25
BitFit	0.10	89.37	94.84	66.96	88.41	92.24	78.70	87.75	91.35	86.20
H-Adapter	1.22	90.13	95.53	68.64	91.91	94.11	84.48	89.95	91.48	88.28
P-Adapter	1.18	90.33	95.61	68.77	92.04	94.29	85.20	89.46	91.54	88.41
LoRA $_{r=8}$	1.33	90.65	94.95	69.82	91.99	93.87	85.20	89.95	91.60	88.50
AdaLoRA	1.27	**90.76**	96.10	71.45	<u>92.23</u>	<u>94.55</u>	88.09	90.69	91.84	89.46
OFT $_{b=16}$	0.79	90.33	96.33	**73.91**	92.10	94.07	87.36	92.16	<u>91.91</u>	89.77
BOFT $_{b=8}^{m=2}$	0.75	90.25	96.44	72.95	92.10	94.23	<u>88.81</u>	92.40	**91.92**	89.89
HRA $_{r=8,\ \lambda=0}$	0.66	<u>90.70</u>	<u>96.45</u>	<u>73.70</u>	91.29	**94.66**	88.45	<u>93.69</u>	91.86	**90.10**
HRA $_{r=8,\ \lambda=10^{-5}}$	0.66	90.43	**96.79**	71.91	91.02	94.44	**89.53**	**94.10**	91.74	<u>90.00</u>
HRA $_{r=8,\ \lambda=\infty}$	0.66	90.52	95.87	70.71	90.71	94.12	87.00	92.59	91.54	89.13

Experiments: Controllable text-to-image generation

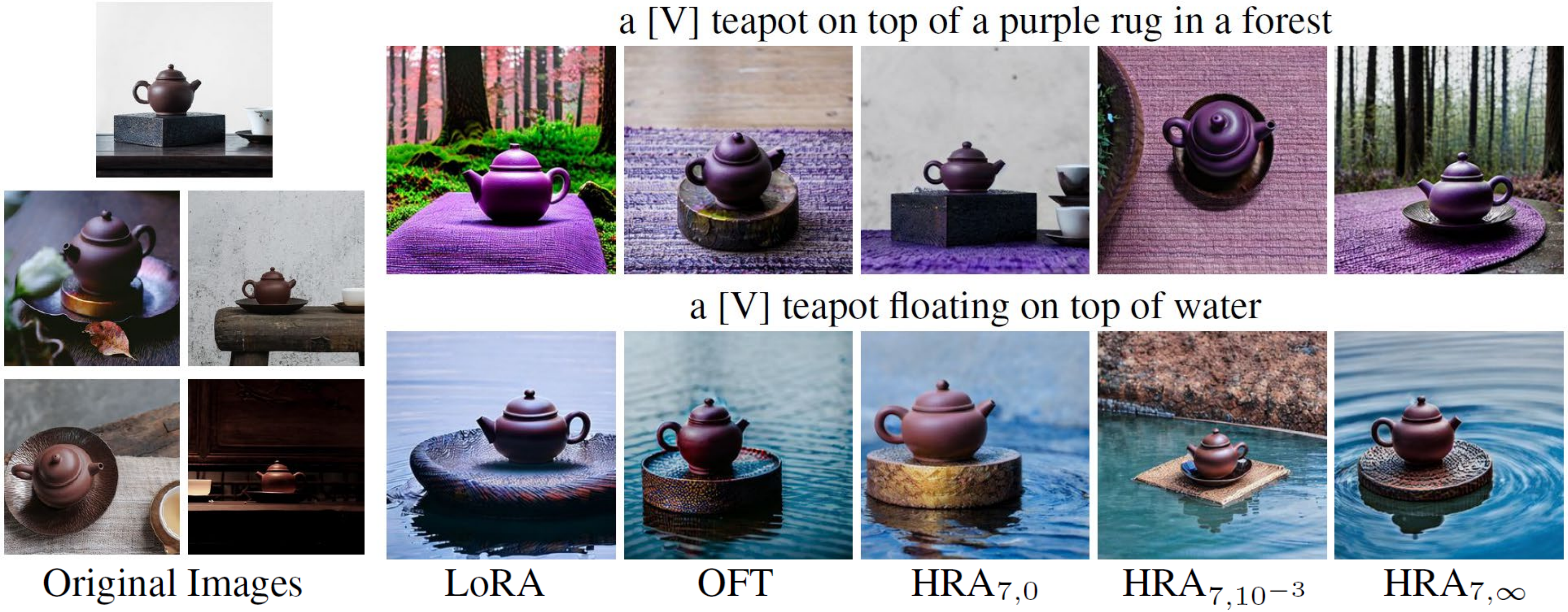


Figure: Qualitative results on subject-driven generation.

Experiments: Controllable text-to-image generation

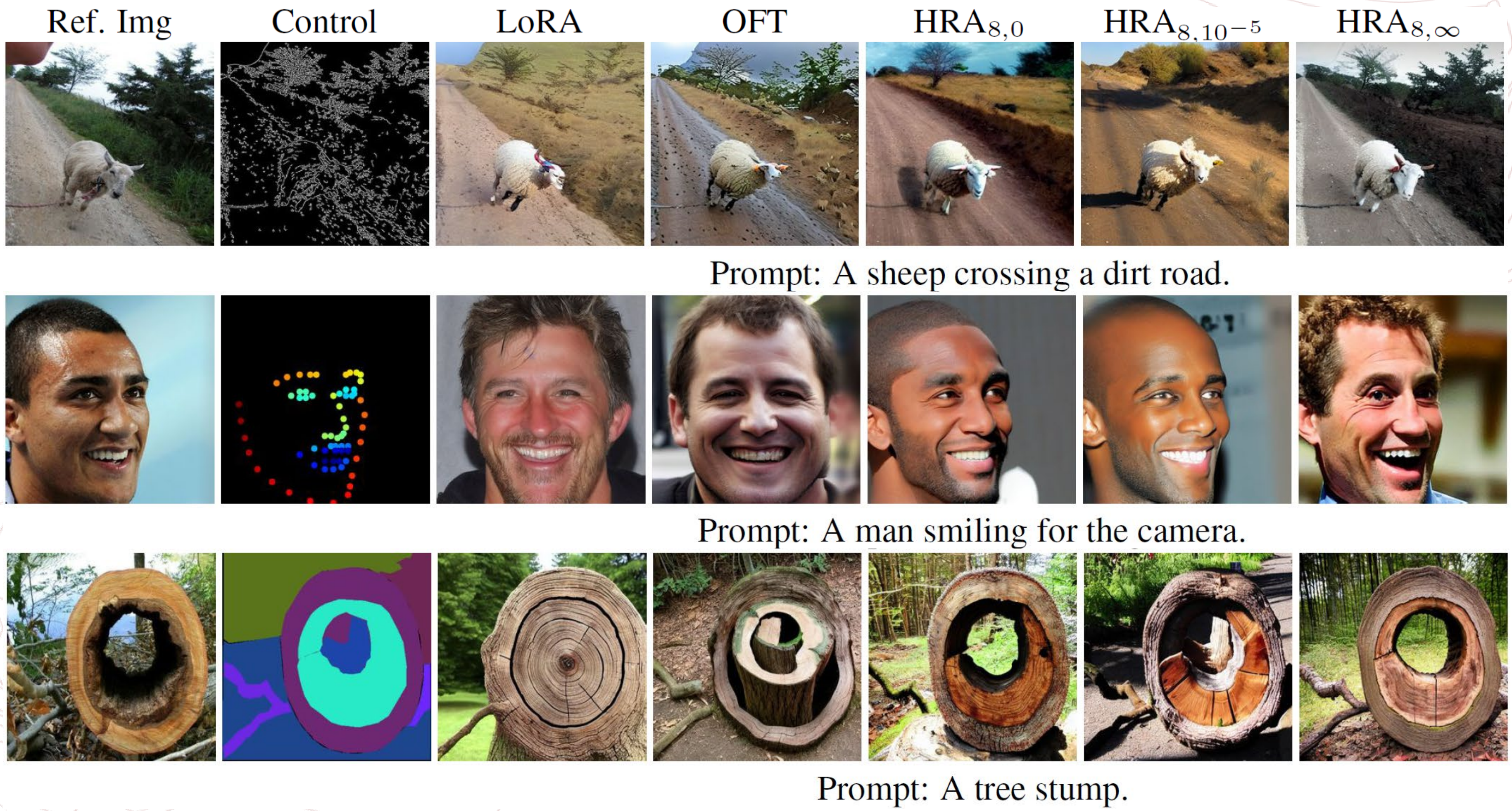


Figure: Qualitative results on controllable generation.

中国人民大學
RENMIN UNIVERSITY OF CHINA

高瓴人工智能学院
Gaoling School of Artificial Intelligence

Thank you for listening!

The code is available at https://github.com/DaShenZi721/HRA and PEFT!