

GradCraft: Elevating Multi-task Recommendations through Holistic Gradient Crafting

USTC & Kuaishou Technology

Yimeng Bai, Yang Zhang, Fuli Feng, Jing Lu, Xiaoxue Zang, Chenyi Lei, Yang Song

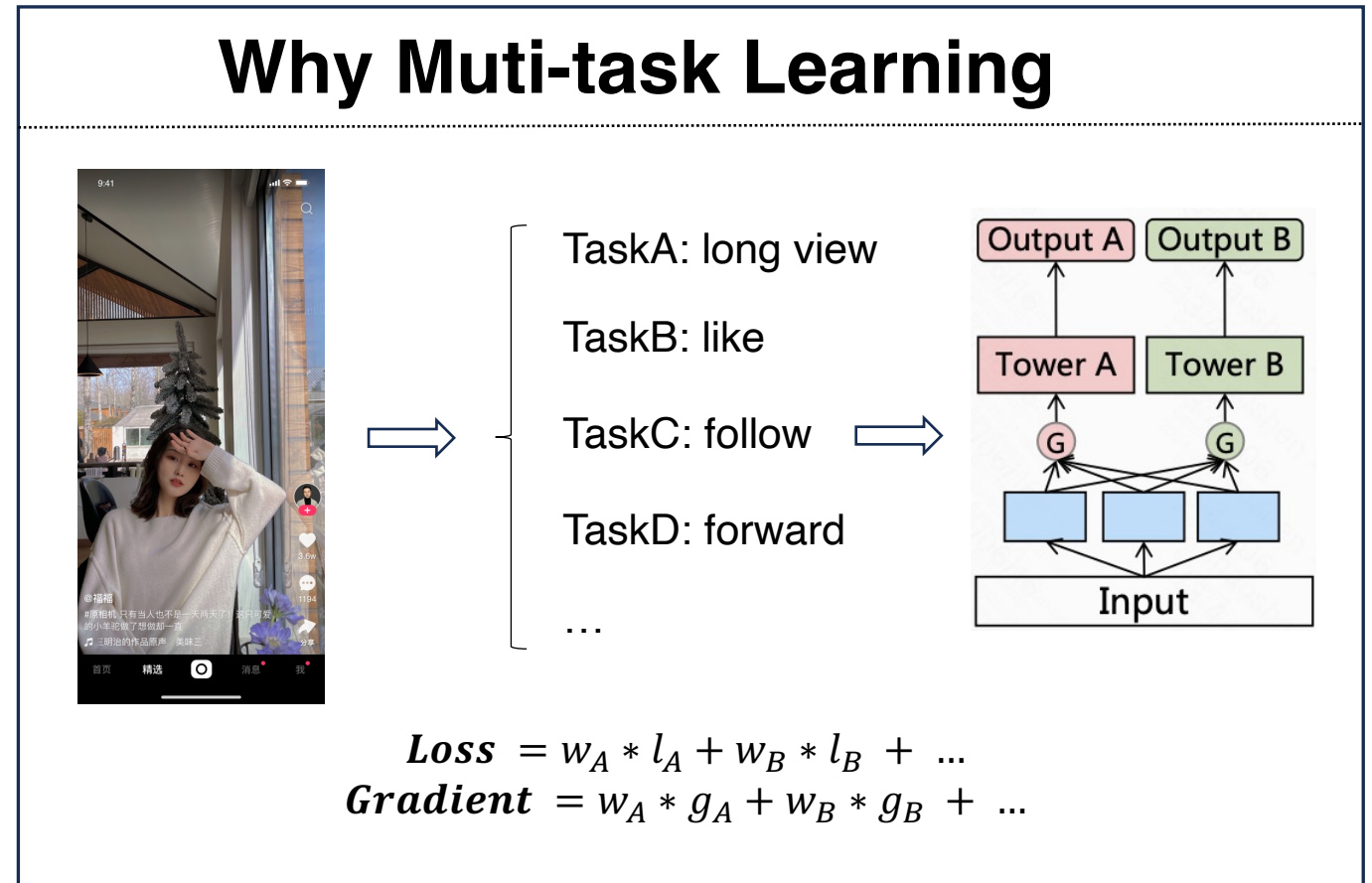
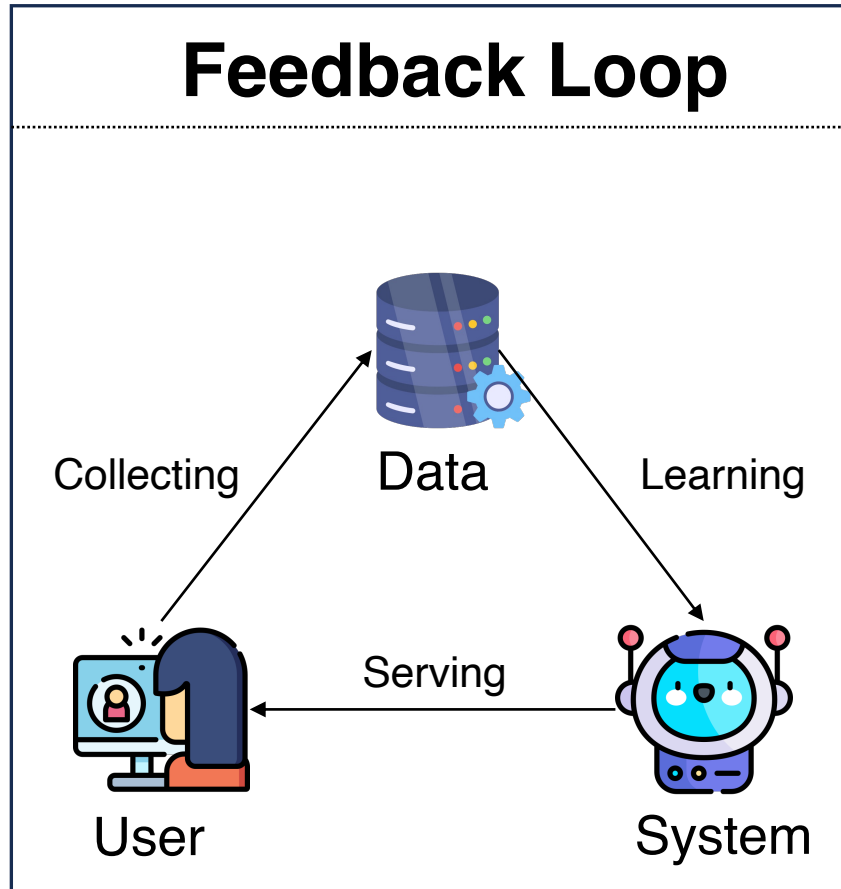


Outline

- Background
- Methodology
- Experiment
- Conclusion

Background

➤ What is Multi-task Recommendation



Background

➤ Challenge in Multi-task Recommendations

Task Heterogeneity



Viewing: long view



Engagement: like

Engagement behavior is **sparser** than viewing behavior

Gradient perspective: **gradient magnitudes** are different.

$$\|g_A\| = 10 \quad \|g_B\| = 1e - 4$$

Task Cardinality

TaskA: long view

TaskB: like

TaskC: follow

...



Gradient perspective: **gradient directions** are different.



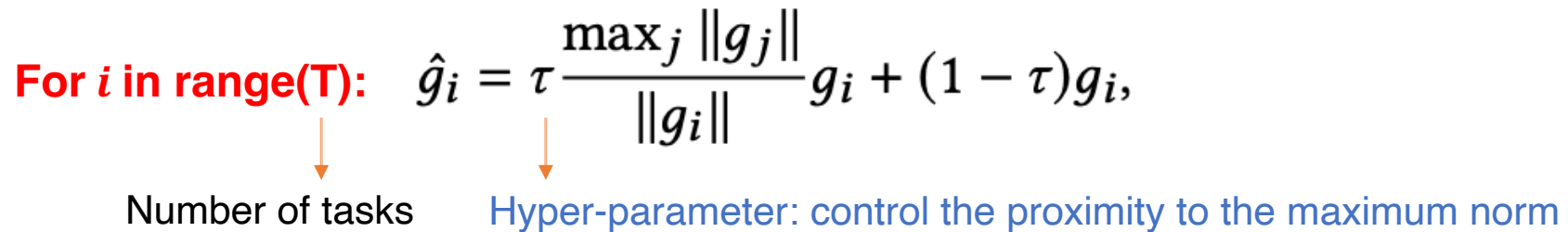
Conflicting task pairs (inner product < 0)

Methodology

➤ Gradient Magnitude Adjustment

- **Key:** ensure an appropriate level of magnitude balance
- **How:** align gradient norm with the maximum norm among tasks

For i in range(T): $\hat{g}_i = \tau \frac{\max_j \|g_j\|}{\|g_i\|} g_i + (1 - \tau) g_i,$



Number of tasks Hyper-parameter: control the proximity to the maximum norm

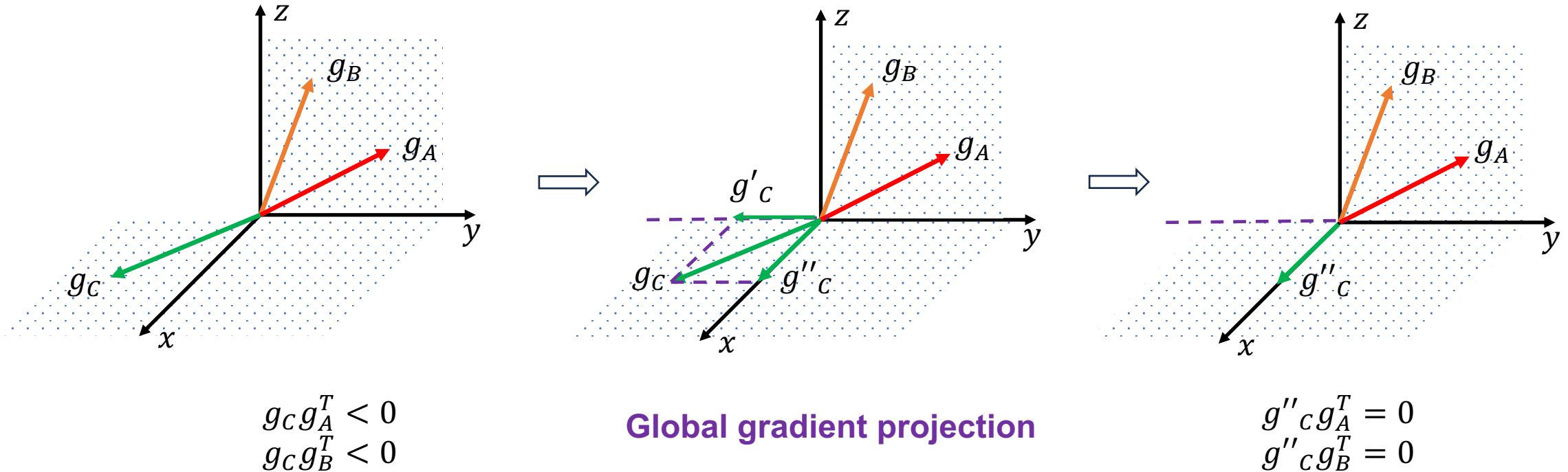
After adjustment: $\max_j \|g_j\| / \|\hat{g}_i\| < 1/\tau$

- Mitigate interference from magnitudes for subsequent manipulation

Methodology

➤ Gradient Direction Deconfliction

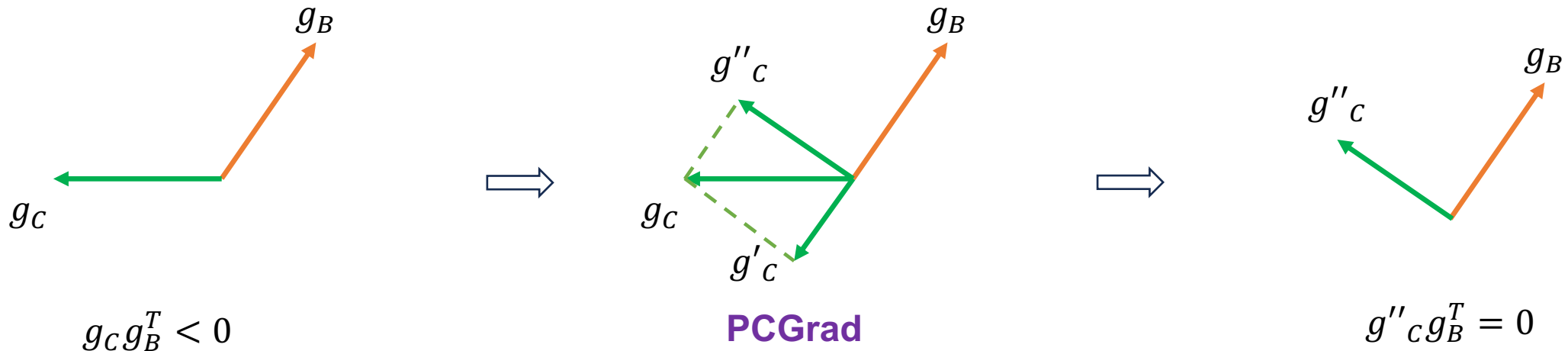
- ❑ **Key:** ensure task gradient does not conflict with other gradients
- ❑ **How:** global gradient projection



Methodology

➤ Gradient Direction Deconfliction

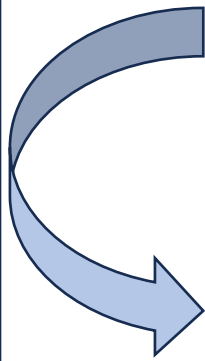
- ❑ For each task, our method **simultaneously** addresses all conflicting tasks
- ❑ Comparison: PCGrad is one vs one, our projection method is one vs all
- ❑ When only **two tasks**: our projection method degrades to PCGrad



Methodology

➤ Gradient Direction Deconfliction

□ Formulation of global gradient projection

✓ Stack all conflicting gradients	$G_i = [\hat{g}_{i_1}, \dots, \hat{g}_{i_n}] \in \mathbb{R}^{n \times d},$	For i in range(T):  Input: \hat{g}_i Output: \tilde{g}_i
✓ Set the projection target	$G_i \tilde{g}_i^T = z,$ $z = [\epsilon \ \hat{g}_i\ \ \hat{g}_{i_1}\ , \dots, \epsilon \ \hat{g}_i\ \ \hat{g}_{i_n}\],$ <small>Hyper-parameter: allow the positive inner product</small>	
✓ Give the projection to the space	$\tilde{g}_i = \hat{g}_i + \sum_{k=1}^n w_k \hat{g}_{i_k} = \hat{g}_i + \mathbf{w}^T G_i,$	
✓ Solve the weight vector	$\mathbf{w} = (G_i G_i^T)^{-1} (-G_i \hat{g}_i^T + z).$	

Methodology

➤ Overall Framework of GradCraft

- ❑ 1. Compute all task losses, $[l_1, \dots, l_T]$
- ❑ 2. Obtain all task gradients, $[g_1, \dots, g_T]$
- ❑ 3. Gradient magnitude adjustment, $[\hat{g}_1, \dots, \hat{g}_T]$
- ❑ 4. Gradient direction deconfliction, $[\tilde{g}_1, \dots, \tilde{g}_T]$
- ❑ 5. Gradient combination, just average as $\frac{1}{T} \sum_{i=1}^T \tilde{g}_i$
- ❑ 6. Gradient update by the optimizer (shared parameters)

Experiment

➤ Offline Experiment Setting

- ❑ **Dataset:** Kuaishou (private), Wechat (public)
- ❑ **Task (Binary for simplicity):**
 - ❑ Viewing behavior: EffectiveView, LongView, CompleteView
 - ❑ Engagement behavior: Like, Follow, Forward
- ❑ **Evaluation:** average value of all tasks' AUC and GAUC
- ❑ **Baseline:**
 - ❑ Simple: Single, EqualWeighting
 - ❑ Other multi-task learning methods like Uncertainty and PCGrad

Experiment

➤ Offline Experiment Result

		Wechat											
Method		Single	EW	UC	DWA	MGDA	PCGrad	PCGrad+	GradVac	CAGrad	IMTL	DBMTL	GradCraft
AUC	EV	0.7641	0.7641	0.7633	0.7646	0.7569	<u>0.7651</u>	0.7644	0.7648	0.7647	0.7629	0.7636	0.7653
	LV	0.8484	0.8484	0.8479	<u>0.8490</u>	0.8429	0.8491	0.8486	0.8489	0.8489	0.8478	0.8479	<u>0.8490</u>
	CV	0.7610	0.7604	0.7596	0.7620	0.7515	0.7614	0.7611	0.7613	0.7614	0.7589	0.7597	<u>0.7616</u>
	Like	0.8661	0.8664	<u>0.8671</u>	0.8656	0.8604	0.8675	0.8668	0.8665	0.8662	0.8669	0.8650	0.8661
	Fol	<u>0.8829</u>	0.8810	0.8763	0.8809	0.8803	0.8825	0.8827	0.8791	0.8801	0.8827	0.8750	0.8888
	For	0.8940	0.9012	0.9006	0.8983	0.8937	0.8968	0.9000	0.8991	0.9003	<u>0.9008</u>	0.8987	0.9001
	AV-A	0.8361	0.8369	0.8358	0.8367	0.8309	0.8371	<u>0.8373</u>	0.8366	0.8369	0.8367	0.8350	0.8385
	RI-A	0.000%	0.091%	-0.038%	0.078%	-0.639%	0.118%	<u>0.135%</u>	0.065%	0.099%	0.056%	-0.129%	0.278%
GAUC	EV	0.6207	0.6209	0.6194	0.6189	0.6055	0.6226	0.6195	0.6218	0.6200	0.6201	0.6178	<u>0.6221</u>
	LV	0.7731	0.7745	0.7740	0.7739	0.7684	<u>0.7754</u>	0.7736	0.7755	0.7743	0.7742	0.7732	0.7751
	CV	0.6499	0.6503	0.6489	0.6499	0.6345	<u>0.6515</u>	0.6493	0.6509	0.6491	0.6488	0.6464	0.6518
	Like	0.6324	0.6382	<u>0.6405</u>	0.6368	0.6328	0.6422	0.6380	0.6384	0.6390	0.6393	0.6385	0.6383
	Fol	0.6847	0.6820	<u>0.6962</u>	0.6915	0.6874	0.6899	0.6870	0.6721	0.6930	0.6894	0.6896	0.7003
	For	0.7012	0.7129	0.7154	0.7141	0.7021	<u>0.7164</u>	0.7140	0.7152	0.7135	0.7144	0.7124	0.7176
	AV-G	0.6770	0.6798	0.6824	0.6809	0.6718	<u>0.6830</u>	0.6802	0.6790	0.6815	0.6810	0.6796	0.6842
	RI-G	0.000%	0.413%	0.791%	0.559%	-0.809%	<u>0.887%</u>	0.472%	0.288%	0.653%	0.589%	0.380%	1.056%

Experiment

➤ Online Experiment on Kuaishou

□ Setting:

- Traffic: 1 week, 15 million users

- Baseline: the SOTA multi-task learning method on our platform

□ Evaluation:

- the average time users spend watching videos (WT)

- the number of effective video viewing records (VV)

- the instances of video sharing (Share)

- **Result:** WT **+0.505%**, VV **+0.950%**, Share **+1.746%**

Experiment

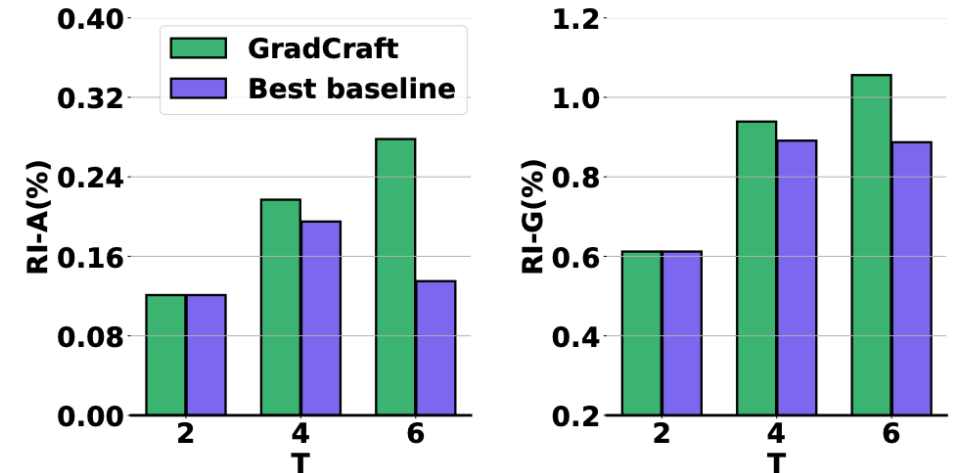
➤ In-depth Analysis

□ Ablation Study

- ✓ ϵ : control the proximity
- ✓ τ : allow the positive inner product
- ✓ Projection: global (one vs all)

	Method	AV-A	RI-A	AV-G	RI-G
	GradCraft	0.8385	0.278%	0.6842	1.056%
$\epsilon = 0$	← GradCraft-fix ϵ	0.8382	0.250%	0.6837	0.981%
$\tau = 1$	← GradCraft-fix τ	0.8365	0.039%	0.6798	0.392%
$\tau = 0$	← GradCraft-ori	0.8370	0.113%	0.6835	0.959%
one vs one	← GradCraft-local	0.8371	0.118%	0.6830	0.887%

□ Effect of Task Number T



- ✓ GradCraft **scales up** effectively with the increasing complexity introduced by a growing number of tasks.

Conclusion

➤ Multi-task Learning in RecSys

- ❑ **Challenge:** task heterogeneity and cardinality
- ❑ **Motivation:** appropriate magnitude balance and global direction balance
- ❑ **Methodology:**
 - ❑ Gradient magnitude adjustment
 - ❑ Gradient direction deconfliction
- ❑ **Offline** experiment on real-world datasets
- ❑ **Online** experiments on Kuaishou platform
- ❑ **Future work:** apply to other domains and improve the efficiency