Beyond Modality Collapse: Representations Blending for Multimodal Dataset Distillation

Xin Zhang[1,2], Ziruo Zhang[1,3], Jiawei Du[1,2], Zuozhu Liu[4], Joey Tianyi Zhou[1,2] ✉

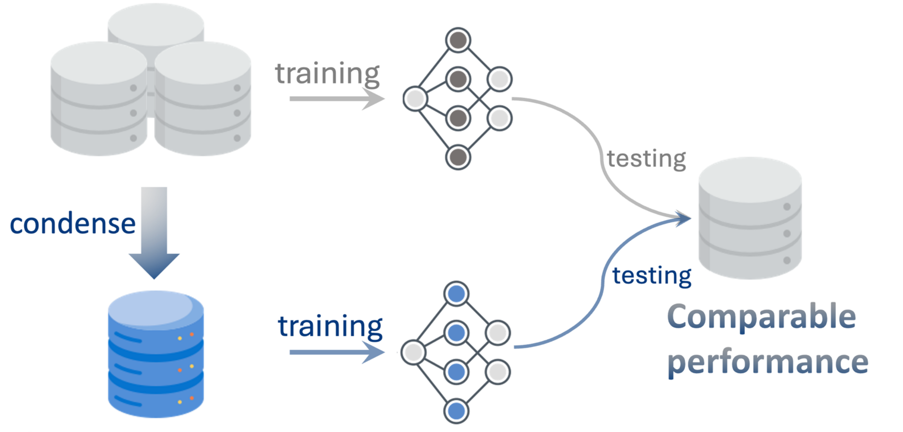
*[1]Centre for Frontier AI Research (CFAR), A*STAR, Singapore, [2]Institute of High Performance Computing (IHPC), A*STAR, Singapore,*
[3]National University of Singapore, Singapore, [4]Zhejiang University, China

paper codes

Centre for Frontier AI Research
A*STAR CFAR

Background

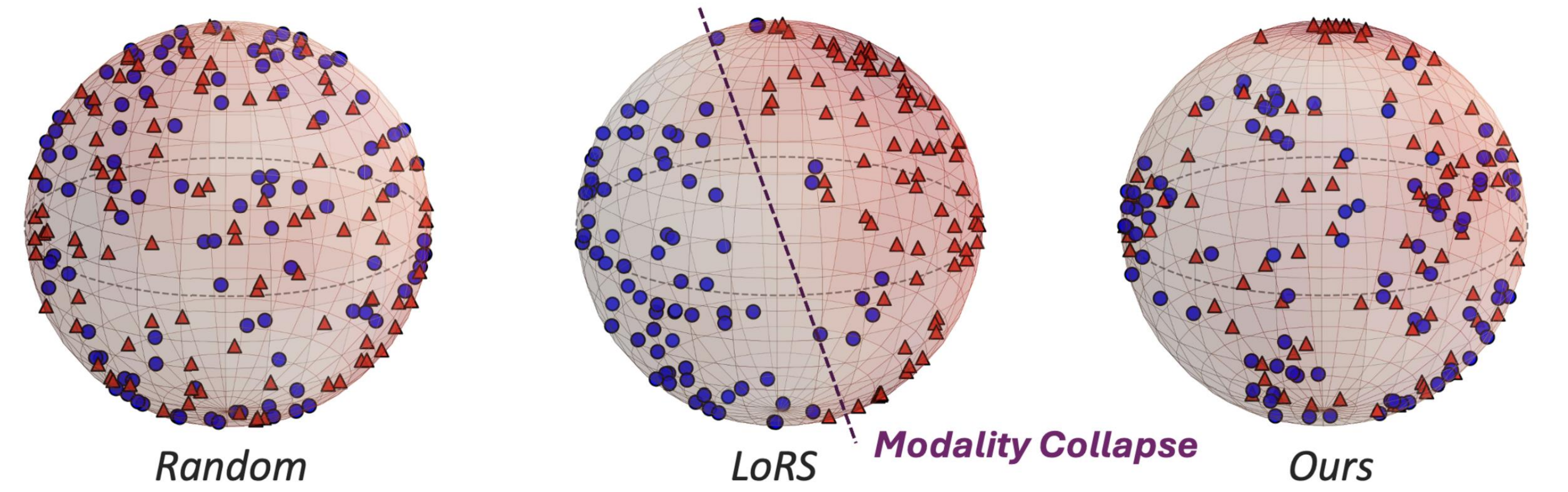
What is Dataset Distillation (DD) ?



Dataset distillation *is to* ***synthesize*** *a* ***tiny and compact dataset*** *from a given* ***real and large dataset****, such that the former can yield a* ***comparable performance*** *as the latter.*

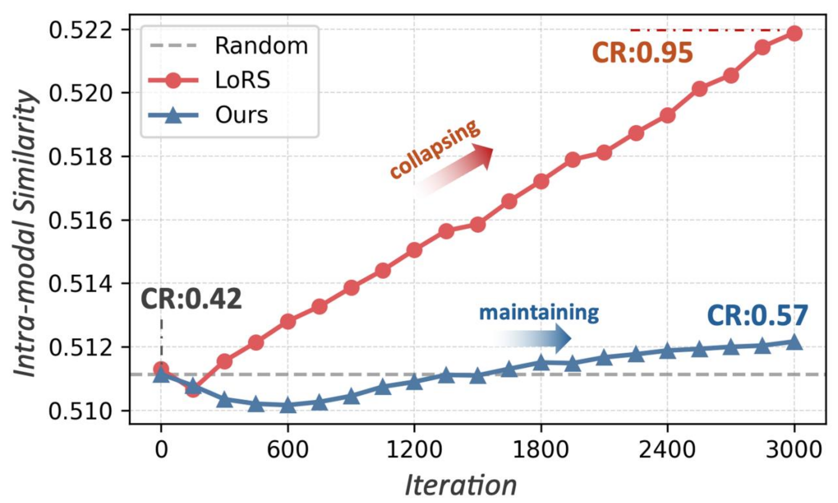
How do existing methods achieve MDD?

❖ *Modality Collapse in MDD*

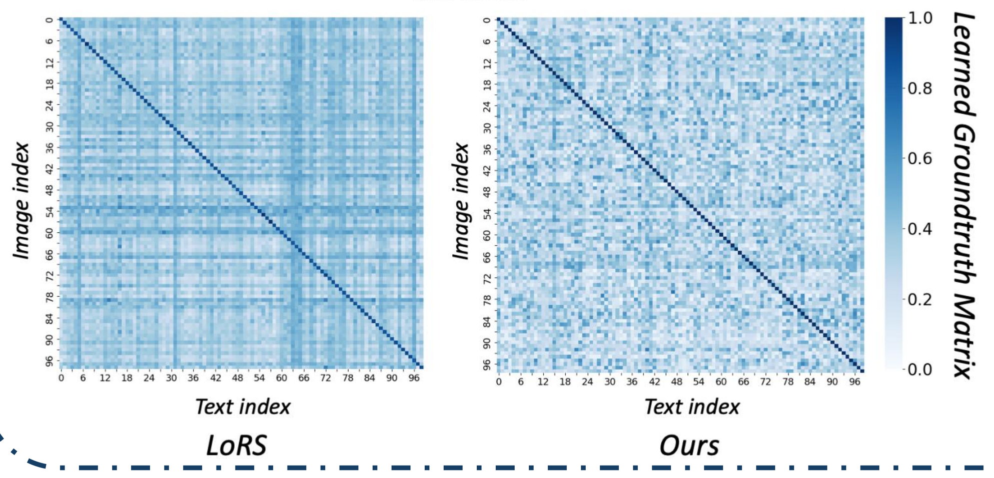


pronounced intra-modality aggregation and ***inter-modality separation***

observable effects:



➢ *The intra-modal similarity consistently increases throughout the distillation process.*



➢ *Feature centralization widens the modality gap, making non-matching pairs indistinguishable.*

Proposed Method

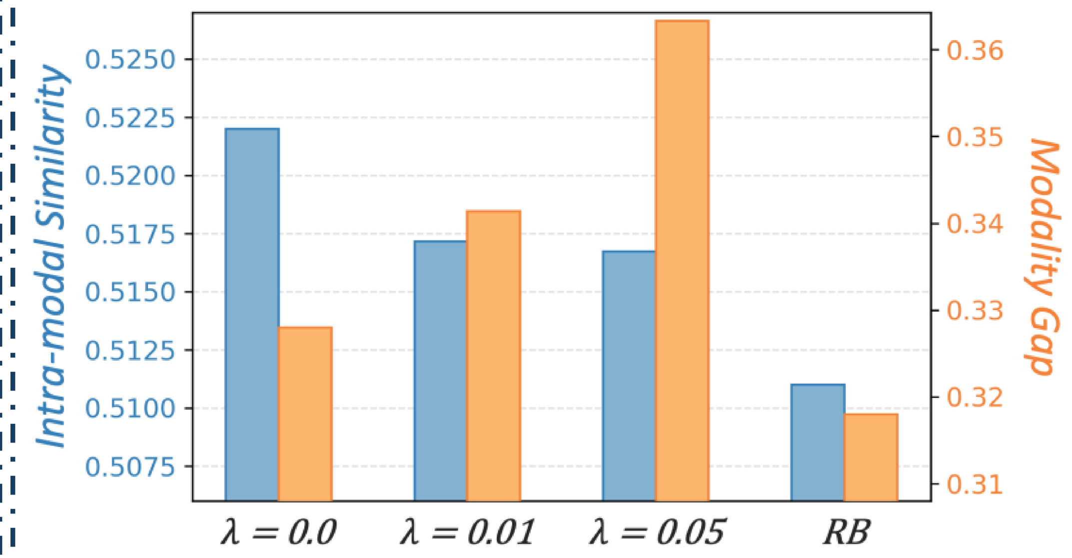
Beyond Modality Collapse

❖ *underlying theoretical cause*

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{\tau})\sim\mathcal{P}}[\mathcal{L}(f_{\boldsymbol{\theta}_\mathcal{S}}(\boldsymbol{x},\boldsymbol{\tau}),\boldsymbol{y})] \quad \text{s.t.} \quad \boldsymbol{\theta}_\mathcal{S} = \arg\min_{\boldsymbol{\theta}} \mathop{\mathbb{E}}_{(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{\tau}})\sim\mathcal{S}}[\mathcal{L}(f_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{\tau}}),\tilde{\boldsymbol{y}})],$$

$$\frac{\partial\mathcal{L}}{\partial\tilde{\boldsymbol{x}}'_n}\frac{\partial\mathcal{L}}{\partial\tilde{\boldsymbol{x}}'_m} \Longleftarrow \frac{w_{nm}w_{mn}}{\gamma^2}[\sigma(\hat{\boldsymbol{y}}_{nm})/t-\tilde{\boldsymbol{y}}_{nm}][\sigma(\hat{\boldsymbol{y}}_{mn})/t-\tilde{\boldsymbol{y}}_{mn}]\tilde{\boldsymbol{\tau}}'^{\top}_m\tilde{\boldsymbol{\tau}}'_n$$

✓ *Mitigating Modality Collapse via Representation Blending*

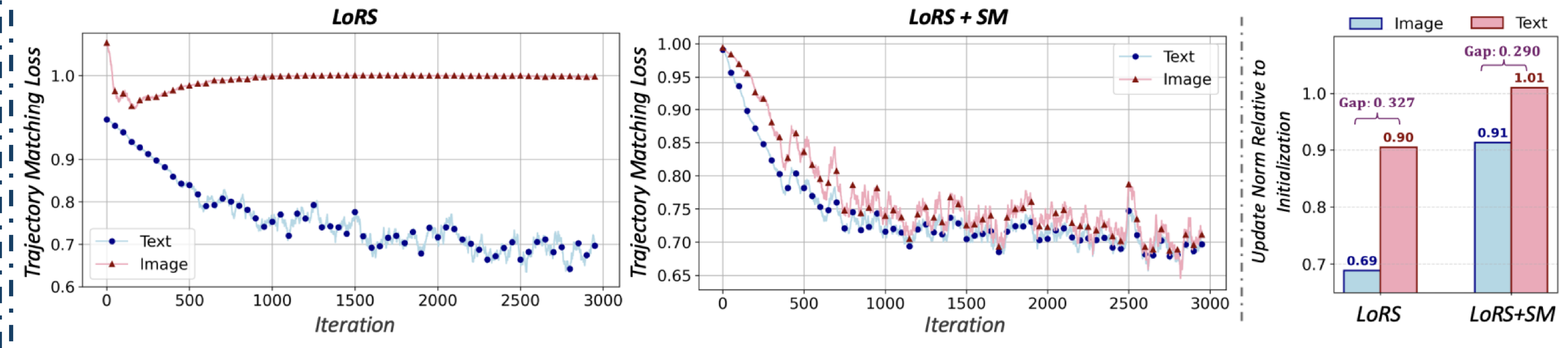


Reduced intra-modal similarity
&
Reduced inter-modal gap

$$\tilde{\boldsymbol{\tau}}'^{\text{+noise}}_m = \text{Normalize}\left(f^{\text{textP}}((1-\lambda)\tilde{\boldsymbol{\tau}}_m+\lambda\vec{\Delta}_m)\right) \Rightarrow \tilde{\boldsymbol{\tau}}'^{\text{blend}}_m = \text{Normalize}\left(f^{\text{textP}}((1-\lambda)\tilde{\boldsymbol{\tau}}_m+\lambda\tilde{\boldsymbol{\tau}}_i)\right)$$

$$\tilde{\boldsymbol{\tau}}'^{\text{+noise}}_n = \text{Normalize}\left(f^{\text{textP}}((1-\lambda)\tilde{\boldsymbol{\tau}}_n+\lambda\vec{\Delta}_n)\right) \quad \tilde{\boldsymbol{\tau}}'^{\text{blend}}_n = \text{Normalize}\left(f^{\text{textP}}((1-\lambda)\tilde{\boldsymbol{\tau}}_n+\lambda\tilde{\boldsymbol{\tau}}_j)\right)$$

✓ *Enhancing Cross-modal Alignment via Symmetric Projection Trajectory Matching*



$$\tilde{\boldsymbol{x}}^*,\tilde{\boldsymbol{\tau}}^*,\tilde{\boldsymbol{y}}^* = \arg\min_{\tilde{\boldsymbol{x}},\tilde{\boldsymbol{\tau}},\tilde{\boldsymbol{y}}}\left(\left\|\boldsymbol{\theta}^{t+T}_{\mathcal{S}_{\text{imgE}}}-\boldsymbol{\theta}^{t+M}_{\mathcal{D}_{\text{imgE}}}\right\|_2^2+\left\|\boldsymbol{\theta}^{t+T}_{\mathcal{S}_{\text{textP}}}-\boldsymbol{\theta}^{t+M}_{\mathcal{D}_{\text{textP}}}\right\|_2^2\right)\Big/\left(\left\|\boldsymbol{\theta}^{t}_{\mathcal{D}_{\text{imgE}}}-\boldsymbol{\theta}^{t+M}_{\mathcal{D}_{\text{imgE}}}\right\|_2^2+\left\|\boldsymbol{\theta}^{t}_{\mathcal{D}_{\text{textP}}}-\boldsymbol{\theta}^{t+M}_{\mathcal{D}_{\text{textP}}}\right\|_2^2\right)$$

$$\Downarrow$$

$$\tilde{\boldsymbol{x}}^*,\tilde{\boldsymbol{\tau}}^*,\tilde{\boldsymbol{y}}^* = \arg\min_{\tilde{\boldsymbol{x}},\tilde{\boldsymbol{\tau}},\tilde{\boldsymbol{y}}}\left(\left\|\boldsymbol{\theta}^{t+T}_{\mathcal{S}_{\text{imgP}}}-\boldsymbol{\theta}^{t+M}_{\mathcal{D}_{\text{imgP}}}\right\|_2^2+\left\|\boldsymbol{\theta}^{t+T}_{\mathcal{S}_{\text{textP}}}-\boldsymbol{\theta}^{t+M}_{\mathcal{D}_{\text{textP}}}\right\|_2^2\right)\Big/\left(\left\|\boldsymbol{\theta}^{t}_{\mathcal{D}_{\text{imgP}}}-\boldsymbol{\theta}^{t+M}_{\mathcal{D}_{\text{imgP}}}\right\|_2^2+\left\|\boldsymbol{\theta}^{t}_{\mathcal{D}_{\text{textP}}}-\boldsymbol{\theta}^{t+M}_{\mathcal{D}_{\text{textP}}}\right\|_2^2\right)$$

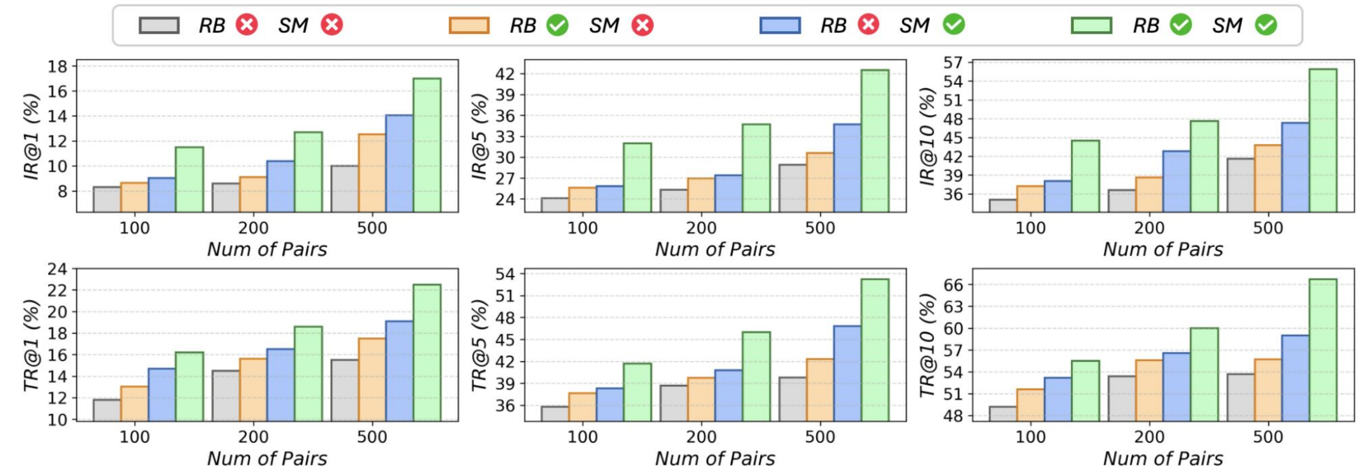
Enhance the image-side distillation

Experiments

Results:

Pairs	Ratio	Metric	Coreset Selection				Dataset Distillation			
			Rand	Herd [50]	K-Cent [16]	Forget [45]	MTT-VL [51]	TESLA-VL [52]	LoRS [52]	Ours
Flickr-30k										
500	1.7%	IR@1	2.4	3.0	3.5	1.8	6.6±0.3	1.1±0.2	10.0±0.2	**17.0**±0.6
		IR@5	10.5	10.0	10.4	9.0	20.2±1.2	7.3±0.4	28.9±0.7	**42.5**±0.5
		IR@10	17.4	17.0	17.3	15.9	30.0±2.1	12.6±0.5	41.6±0.6	**55.9**±0.6
		TR@1	5.2	5.1	4.9	3.6	13.3±0.6	5.1±0.2	15.5±0.7	**22.5**±0.4
		TR@5	18.3	16.4	16.4	12.3	32.8±1.8	15.3±0.5	39.8±0.4	**53.2**±0.3
		TR@10	25.7	24.3	23.3	19.3	46.8±0.8	23.8±0.3	53.7±0.3	**66.7**±0.3
COCO										
500	4.4‰	IR@1	1.1	1.7	1.1	0.8	2.5±0.5	0.8±0.2	2.8±0.2	**6.2**±0.1
		IR@5	5.0	5.3	6.3	5.8	8.9±0.7	3.6±0.6	9.9±0.5	**19.9**±0.3
		IR@10	8.7	9.9	10.5	8.2	15.8±1.5	6.7±0.9	16.5±0.7	**30.6**±0.1
		TR@1	1.9	1.9	2.5	2.1	5.0±0.4	1.7±0.4	5.3±0.5	**7.0**±0.2
		TR@5	7.5	7.8	8.7	8.2	17.2±1.3	5.9±0.8	18.3±1.5	**22.0**±0.3
		TR@10	12.5	13.7	14.3	13.0	26.0±1.9	10.2±1.0	27.9±1.4	**32.9**±0.6

Comparison with SOTAs on Flickr-30k & COCO



The ablation study of RepBlend

Methods	ImageNet-10 Classification		TextCaps Retrieval					
	ACC@1	ACC@5	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
LoRS [55]	21.4	74.4	1.7	5.1	8.4	0.4	1.7	3.1
Ours	**27.6**	**76.2**	**3.1**	**9.4**	**14.5**	**1.9**	**6.2**	**10.3**

Zero-Shot Generalization

Methods	LoRS [52]	Ours
(IR@1, TR@1) (%)	(8.3, 11.8)	**(11.5, 16.2)**
Buffer		
Speed (min/traj)	70	**40**
Memory (GB/traj)	1.63	**0.73**
Distillation		
Speed (s/iter)	11.5	**1.71**
Peak GPU VRAM (GB)	21.78	**10.17**

Better performance & higher speed

Visualization

References:

Beyond Modality Collapse: Representations Blending for Multimodal Dataset Distillation. ***Zhang, Xin*** *and Zhang, Ziruo, and Du, Jiawei and Liu, Zuozhu and Zhou, Joey Tianyi. NeurIPS 2025.*

Low-Rank Similarity Mining for Multimodal Dataset Distillation. Xu, Yue and Lin, Zhilin and Qiu, Yusong and Lu, Cewu and Li, Yong-Lu. ICML 2024