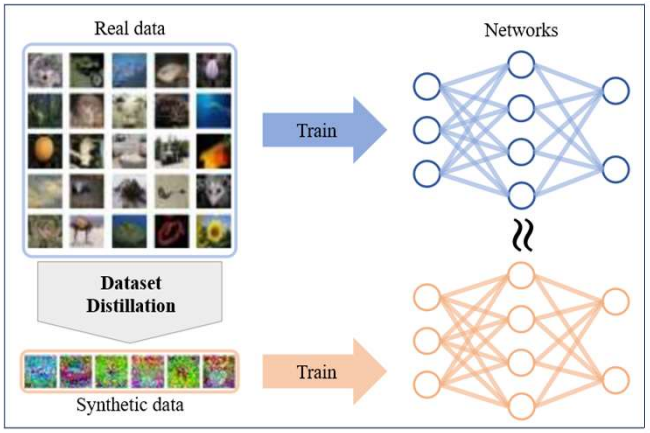
Multimodal Dataset Distillation via Phased Teacher Models

Shengbin Guo* · Hang Zhao* · Senqiao Yang · Chenyang Jiang · Yuhang Cheng · Xiangru Peng · Rui Shao · Zhuotao Tian†
Harbin Institute of Technology, Shenzhen, The Chinese University of Hong Kong

Github: https://github.com/Previsior/PTM-ST
Arxiv: https://arxiv.org/abs/2603.25388
Huggingface: https://huggingface.co/datasets/Previsior22/PTM-ST

Motivation

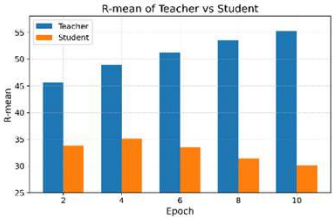


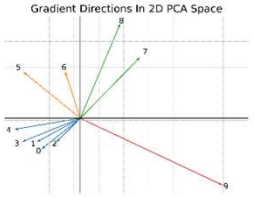
Dataset Distillation technology aims to synthesize a small-scale, high-information-density core data set to replace the huge original data set for efficient training.

Match Training Trajectory method optimizing the synthetic dataset makes the training trajectory of the model on the synthetic dataset approximate the training trajectory on the real dataset.

$$\tau = \left\{ \theta_0, \theta_1, \cdots, \underbrace{\theta_T, \cdots, \theta_{T^+}}_{\text{matching range}}, \cdots, \theta_n \right\}$$

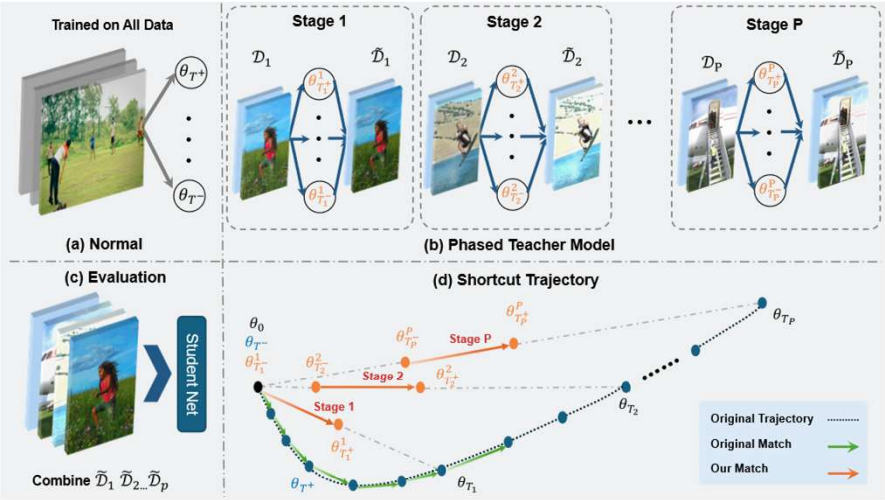
τ is the trajectory to match.
We will specify a matching range.





Simply increasing T^+ will cause the distillation effect to decrease. We found that this is due to the different knowledge contained in the trajectories at different stages, and the gradients are also quite different.

Method



To this end, We proposed PTM-ST. We divided the distillation into several stages, and each stage used different interpolation trajectories as guidance, which effectively alleviated the problem of stage gaps.

Algorithm 1 Distillation: Phased Teacher Model with Shortcut-Trajectory

Input: Real dataset $\mathcal{X}, \mathcal{Y}$. Size, iteration, matching range, interpolation endpoint of each subset: $\{(N_p, I_p, T_p^-, T_p^+, t_p)\}_{p=1}^{P}$.Decay parameter α,learning rate γ.Synthesis steps t, expert epochs ΔT.
Output: A series of synthetic datasets: $\tilde{\mathcal{D}}_1, \tilde{\mathcal{D}}_2, \ldots, \tilde{\mathcal{D}}_P$.

Generating synthetic subsets:
for $p = 1, 2, \ldots, P$ **do**
  Initialize $\tilde{\mathcal{X}}_p, \tilde{\mathcal{Y}}_p$ by real data, $\tilde{S}_p = \mathbb{I}_{N_p}$ (identity matrix). $\hat{\mathcal{D}}_p^0 = (\tilde{\mathcal{X}}_p, \tilde{\mathcal{Y}}_p, \tilde{S}_p)$
  Calculate Shortcut-Trajectory $\{\theta^p_{T_p^-}, \ldots, \theta^p_{T_p^+}\}$ based on θ_0 and θ_{t_p}
  for $i = 1, 2, \ldots, I_p$ **do**
    Sample an initial network parameter θ_T^p in $\{\theta^p_{T_p^-}, \ldots, \theta^p_{T_p^+}\}$
    Train the network for t steps on $\tilde{\mathcal{D}}_p$ to $\tilde{\theta}_T^p$
    Compute MTT loss $\mathcal{L}_{\text{MTT}} = \|\tilde{\theta}_T^p - \theta^p_{T+\Delta T}\|^2 / \|\theta_T^p - \theta^p_{T+\Delta T}\|^2$
    Gradient descent: $\tilde{\mathcal{X}}_p \leftarrow \tilde{\mathcal{X}}_p - \gamma_{\mathcal{X}} \nabla_{\tilde{\mathcal{X}}_p} \mathcal{L}_{\text{MTT}}$, $\tilde{\mathcal{Y}}_p \leftarrow \tilde{\mathcal{Y}}_p - \gamma_{\mathcal{Y}} \nabla_{\tilde{\mathcal{Y}}_p} \mathcal{L}_{\text{MTT}}$, $\tilde{S}_p \leftarrow \tilde{S}_p - \gamma_S \nabla_{\tilde{S}_p} \mathcal{L}_{\text{MTT}}$
    EMA: $\hat{\mathcal{D}}_p^i \leftarrow (\tilde{\mathcal{X}}_p, \tilde{\mathcal{Y}}_p, \tilde{S}_p)$, $\hat{\mathcal{D}}_p^i \leftarrow \alpha \hat{\mathcal{D}}_p^{i-1} + (1-\alpha)\hat{\mathcal{D}}_p^i$
  end for
  $\tilde{\mathcal{D}}_p \leftarrow \hat{\mathcal{D}}_p^{I_p}$
end for

Result

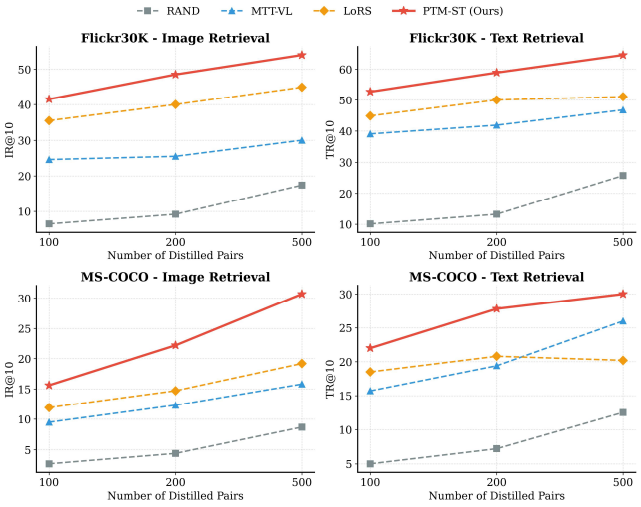


Table 3: Results on LLaVA-cc3m dataset. The metrics for training model on the full dataset are IR@1=9.3, IR@5=25.9, IR@10=36.5; TR@1=9.8, TR@5=26.4, TR@10=37.3.

Pairs	LoRS						Ours					
	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
100	1.2	4.6	7.7	1.7	6.9	11.4	2.3	8.2	13.2	2.9	10.0	15.9
200	1.4	5.3	8.7	2.4	8.5	13.6	2.7	9.7	15.8	3.7	11.9	18.4
500	1.7	6.2	10.1	2.5	8.7	13.8	3.3	11.4	17.9	4.1	13.2	19.9

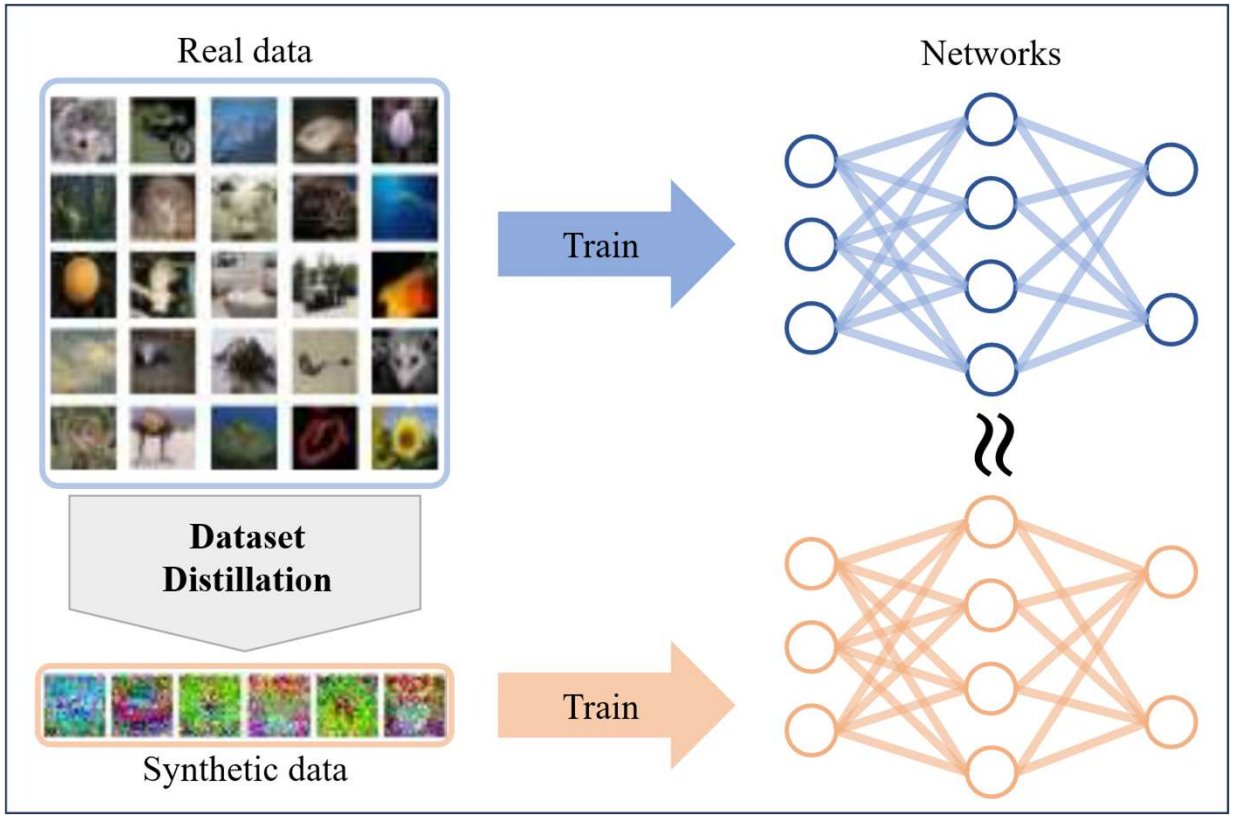
Table 4: Various ablation studies with 500 pairs on Flickr30k and COCO.

No.	Model	Flickr IR@K			Flickr TR@K			COCO IR@K			COCO TR@K		
		1	5	10	1	5	10	1	5	10	1	5	10
(1)	BASE (N/A)	12.2	33.0	45.7	16.2	39.4	54.0	3.4	11.7	19.1	4.2	13.8	20.8
(2)	EMA	12.9	33.7	46.3	16.2	40.6	54.3	3.5	12.3	19.8	4.0	13.6	20.8
(3)	PTM	13.4	35.2	48.1	19.6	43.2	55.5	4.2	14.2	22.6	4.8	15.1	23.4
(4)	ST	14.2	37.8	50.8	19.5	45.1	59.3	4.9	15.9	24.7	4.8	15.1	23.8
(5)	PTM + EMA	14.3	37.7	50.7	18.8	45.0	58.5	4.5	14.8	23.1	4.8	15.7	24.3
(6)	ST + EMA	15.4	38.3	51.4	19.7	46.7	60.3	5.2	17.0	26.3	5.0	15.8	24.0
(7)	PTM + ST	15.4	38.8	52.2	22.3	50.5	64.6	6.3	20.1	30.2	6.2	19.9	29.8
(8)	OURS(PTM + ST + EMA)	**15.5**	**39.6**	**53.6**	**22.9**	**51.6**	**64.9**	**6.6**	**20.5**	**30.7**	**6.9**	**20.1**	**30.0**

The experimental results prove the effectiveness of our method, and the effectiveness of each module is verified through ablation experiments.

Multimodal Dataset Distillation via Phased Teacher Models

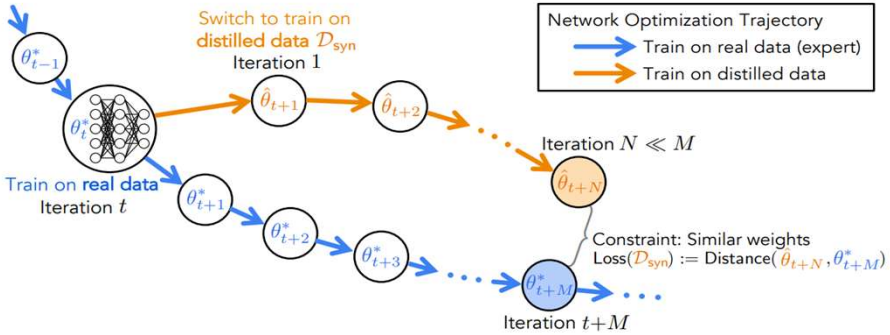
Motivation - background



- Dataset Distillation technology aims to synthesize a small-scale, high-information-density core data set to replace the huge original data set for efficient training.
- Match Training Trajectory method optimizing the synthetic dataset makes the training trajectory of the model on the synthetic dataset approximate the training trajectory on the real dataset.

Multimodal Dataset Distillation via Phased Teacher Models

Motivation - problem

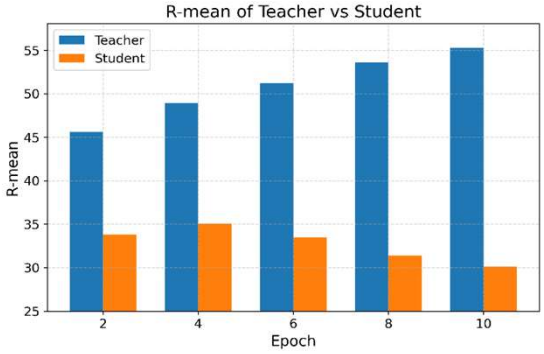


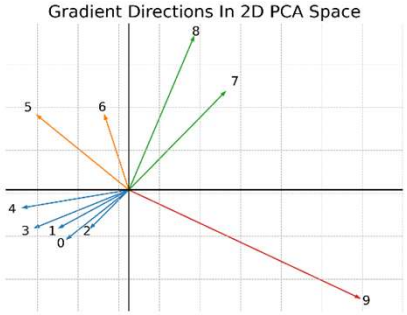
$$\tau = \left\{ \theta_0, \theta_1, \cdots, \underbrace{\theta_{T^-}, \cdots, \theta_{T^+}}_{\textbf{matching range}}, \cdots, \theta_n \right\}$$

- Match Training Trajectory method optimizing the synthetic dataset makes the training trajectory of the model on the synthetic dataset approximate the training trajectory on the real dataset.

$$\mathcal{L} = \frac{\|\hat{\theta}_{t+N} - \theta^*_{t+M}\|_2^2}{\|\theta^*_t - \theta^*_{t+M}\|_2^2},$$

- τ is the trajectory to match.
  MTT will specify a matching range T^-~T^+.

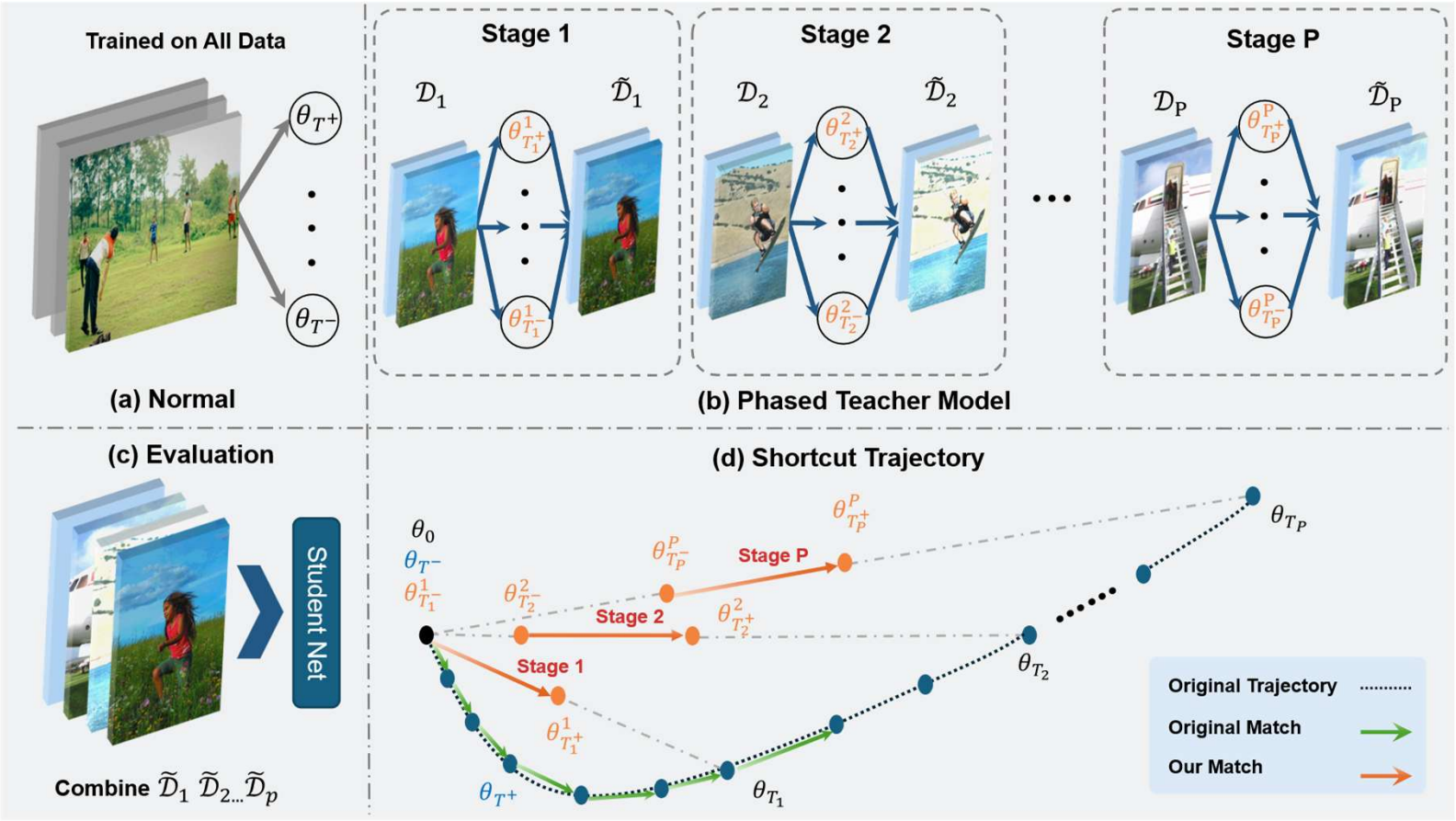




- Simply increasing T^+ will cause the distillation effect to decrease. We found that this is due to the different knowledge contained in the trajectories at different stages, and the gradients are also quite different.

Multimodal Dataset Distillation via Phased Teacher Models

Method



To this end,
We proposed PTM-ST.
We divided the distillation into several stages, and each stage used different interpolation trajectories as guidance, which effectively alleviated the problem of stage gaps.

Multimodal Dataset Distillation via Phased Teacher Models

Algorithm

Algorithm 1 Distillation: Phased Teacher Model with Shortcut-Trajectory

Input: Real dataset $\mathcal{X}, \mathcal{Y}$. Size, iteration, matching range, interpolation endpoint of each subset: $\{(N_p, I_p, T_p^-, T_p^+, t_p)\}_{p=1}^{P}$.Decay parameter α,learning rate γ.Synthesis steps t, expert epochs ΔT.

Output: A series of synthetic datasets: $\tilde{\mathcal{D}}_1, \tilde{\mathcal{D}}_2, \ldots, \tilde{\mathcal{D}}_P$.

Generating synthetic subsets:
for $p = 1, 2, \ldots, P$ **do**
  Initialize $\tilde{\mathcal{X}}_p, \tilde{\mathcal{Y}}_p$ by real data, $\tilde{S}_p = \mathbb{I}_{N_p}$ (identity matrix). $\hat{\mathcal{D}}_p^0 = (\tilde{\mathcal{X}}_p, \tilde{\mathcal{Y}}_p, \tilde{S}_p)$
  Calculate Shortcut-Trajectory $\{\theta^p_{T_p^-}, \ldots, \theta^p_{T_p^+}\}$ based on θ_0 and θ_{t_p}
  for $i = 1, 2, \ldots, I_p$ **do**
    Sample an initial network parameter θ_T^p in $\{\theta^p_{T_p^-}, \ldots, \theta^p_{T_p^+}\}$
    Train the network for t steps on $\tilde{\mathcal{D}}_p$ to $\tilde{\theta}_T^p$
    Compute MTT loss $\mathcal{L}_{\text{MTT}} = \|\tilde{\theta}_T^p - \theta^p_{T+\Delta T}\|^2 / \|\theta_T^p - \theta^p_{T+\Delta T}\|^2$
    Gradient descent: $\tilde{\mathcal{X}}_p \leftarrow \tilde{\mathcal{X}}_p - \gamma_{\mathcal{X}} \nabla_{\tilde{\mathcal{X}}_p} \mathcal{L}_{\text{MTT}}, \tilde{\mathcal{Y}}_p \leftarrow \tilde{\mathcal{Y}}_p - \gamma_{\mathcal{Y}} \nabla_{\tilde{\mathcal{Y}}_p} \mathcal{L}_{\text{MTT}}$, $\tilde{S}_p \leftarrow \tilde{S}_p - \gamma_S \nabla_{\tilde{S}_p} \mathcal{L}_{\text{MTT}}$
    EMA: $\tilde{\mathcal{D}}_p^i \leftarrow (\tilde{\mathcal{X}}_p, \tilde{\mathcal{Y}}_p, \tilde{S}_p)$, $\hat{\mathcal{D}}_p^i \leftarrow \alpha \hat{\mathcal{D}}_p^{i-1} + (1-\alpha)\tilde{\mathcal{D}}_p^i$
  end for
  $\tilde{\mathcal{D}}_p \leftarrow \hat{\mathcal{D}}_p^{I_p}$
end for

Multimodal Dataset Distillation via Phased Teacher Models

Result

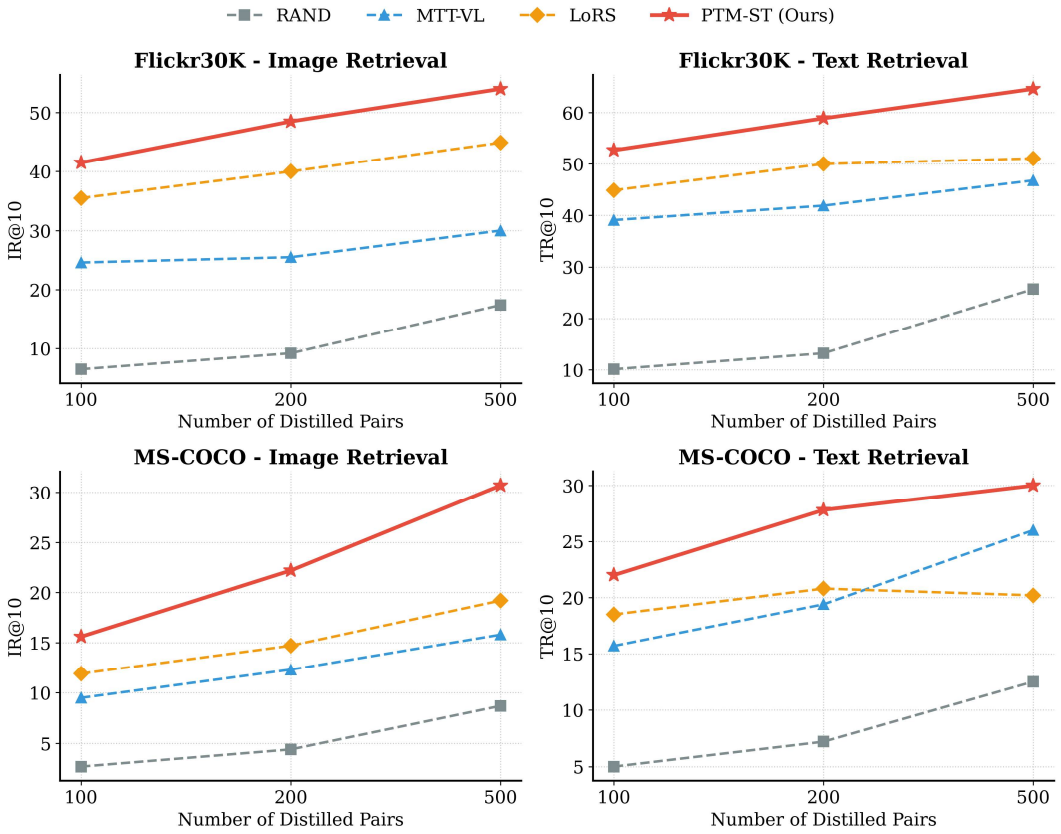


Table 3: Results on LLaVA-cc3m dataset. The metrics for training model on the full dataset are IR@1=9.3, IR@5=25.9, IR@10=36.5; TR@1=9.8, TR@5=26.4, TR@10=37.3.

Pairs	LoRS						Ours					
	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
100	1.2	4.6	7.7	1.7	6.9	11.4	2.3	8.2	13.2	2.9	10.0	15.9
200	1.4	5.3	8.7	2.4	8.5	13.6	2.7	9.7	15.8	3.7	11.9	18.4
500	1.7	6.2	10.1	2.5	8.7	13.8	3.3	11.4	17.9	4.1	13.2	19.9

Table 4: Various ablation studies with 500 pairs on Flickr30k and COCO.

No.	Model	Flickr IR@K			Flickr TR@K			COCO IR@K			COCO TR@K		
		1	5	10	1	5	10	1	5	10	1	5	10
(1)	BASE (N/A)	12.2	33.0	45.7	16.2	39.4	54.0	3.4	11.7	19.1	4.2	13.8	20.8
(2)	EMA	12.9	33.7	46.3	16.2	40.6	54.3	3.5	12.3	19.8	4.0	13.6	20.8
(3)	PTM	13.4	35.2	48.1	19.6	43.2	55.5	4.2	14.2	22.6	4.8	15.1	23.4
(4)	ST	14.2	37.8	50.8	19.5	45.1	59.3	4.9	15.9	24.7	4.8	15.1	23.8
(5)	PTM + EMA	14.3	37.7	50.7	18.8	45.0	58.5	4.5	14.8	23.1	4.8	15.7	24.3
(6)	ST + EMA	15.4	38.3	51.4	19.7	46.7	60.3	5.2	17.0	26.3	5.0	15.8	24.0
(7)	PTM + ST	15.4	38.8	52.2	22.3	50.5	64.6	6.3	20.1	30.2	6.2	19.9	29.8
(8)	OURS(PTM + ST + EMA)	**15.5**	**39.6**	**53.6**	**22.9**	**51.6**	**64.9**	**6.6**	**20.5**	**30.7**	**6.9**	**20.1**	**30.0**

The experimental results prove the effectiveness of our method, and the effectiveness of each module is verified through ablation experiments.