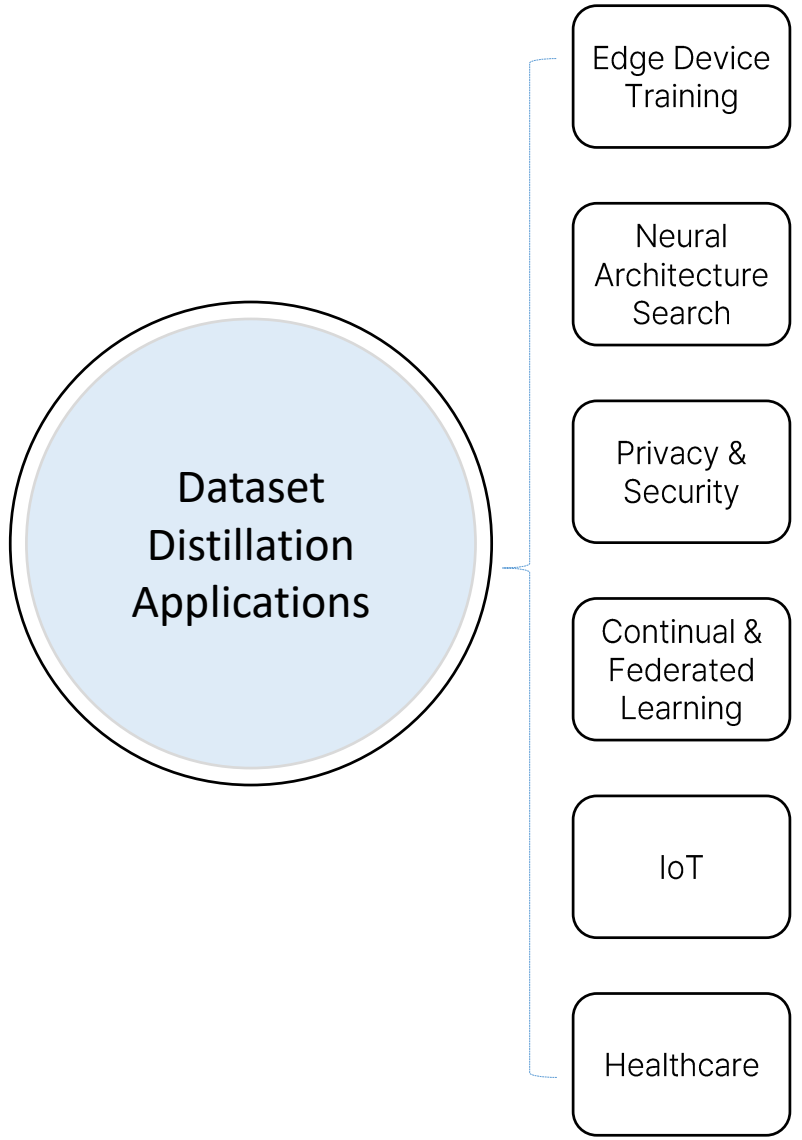
Multimodal Distribution Matching for Vision-Language Dataset Distillation

Jongoh Jeong*, Hoyong Kwon*, Minseok Kim* and Kuk-Jin Yoon

Visual Intelligence Lab., KAIST

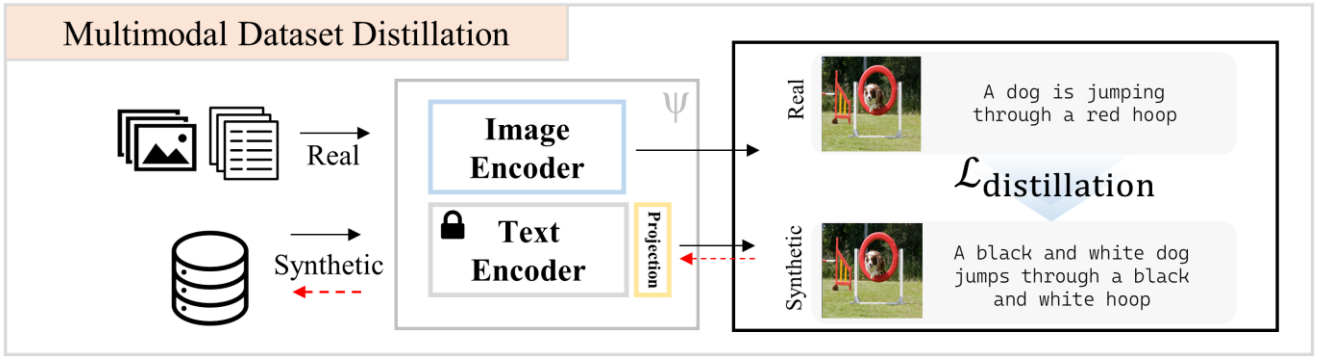
Background: Dataset Distillation

- **Dataset distillation** (DD) compresses a large training set into a small set of synthetic samples that retain its learning utility.
  - **Why compress the dataset?**
    - Modern training pipelines are often limited by data scale, storage, and repetitive training cost
    - DD relieves this cost by building a compact yet highly informative core, enabling efficient learning under tight budgets
  - **What should a distilled dataset preserve?**
    - The most transferable training signals of the original dataset
    - Strong generalization performance, despite using only a tiny number of synthetic samples
  - **Practical significance**
    - Allows repeated training feasible for resource-constrained and iterative settings
    - Broadly useful for applications including edge device deployment, NAS, privacy-sensitive learning, federated learning, IoT, and healthcare



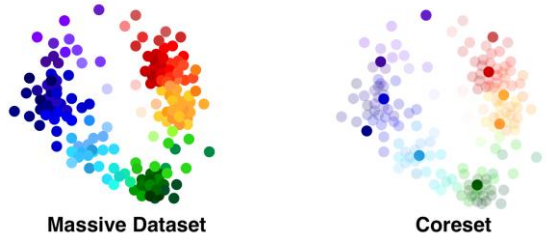
Background: Multimodal Dataset Distillation

- **Multimodal dataset distillation** (MDD) compresses large-scale image-text dataset into compact yet representative synthetic pairs.
  - **Aim**
    - Learn a small set of synthetic image-text pairs
    - Retain the training utility of the full real multimodal dataset
  - **Why compress multimodal data?**
    - Reduce storage and training cost for vision-language learning
    - Enable faster experimentation and deployment with compact synthetic data
  - **What must be preserved?**
    - Intra-modal structure within image and text embeddings
    - Cross-modal alignment between paired images and captions

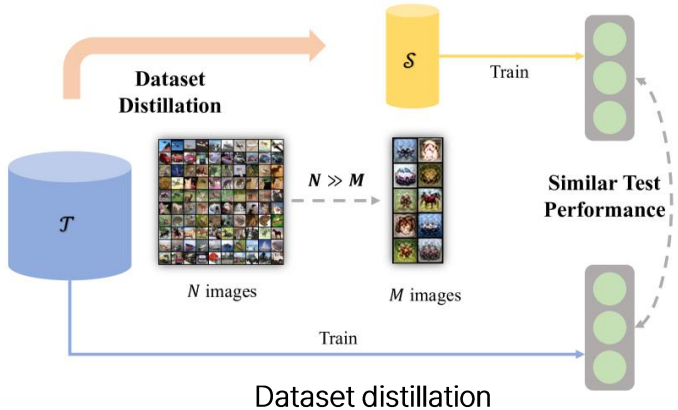


Background: Why Prior Methods Fall Short

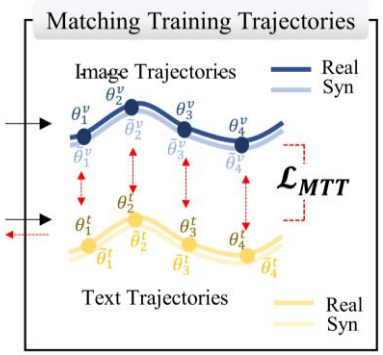
- **Previous approaches**
  - **Coreset Selection**
    - Selects a subset of real image-text pairs that approximately covers the dataset feature space
    - Efficient and simple, yet essentially a re-sampling strategy
    - Coverage-based heuristics often miss joint semantic modes and weakly preserve cross-modal structure
  - **Trajectory / Similarity-based Supervision**
    - Prior MDD methods replay training trajectories or preserve low-rank similarity structure
    - While this improves fidelity to the training dynamics, it relies on indirect supervision from particular experts rather than directly matching the multimodal data distribution itself
- **Limitations**
  - **High computational overhead**
    - Trajectory-based distillation requires nested gradient and bi-level optimization, yielding high compute and memory cost → inefficient distillation
  - **Weak Cross-Architecture Generalization**
    - Synthetic data becomes specialized to the source encoder's geometry → Transfer to different image-text geometry is limited



Coreset Selection

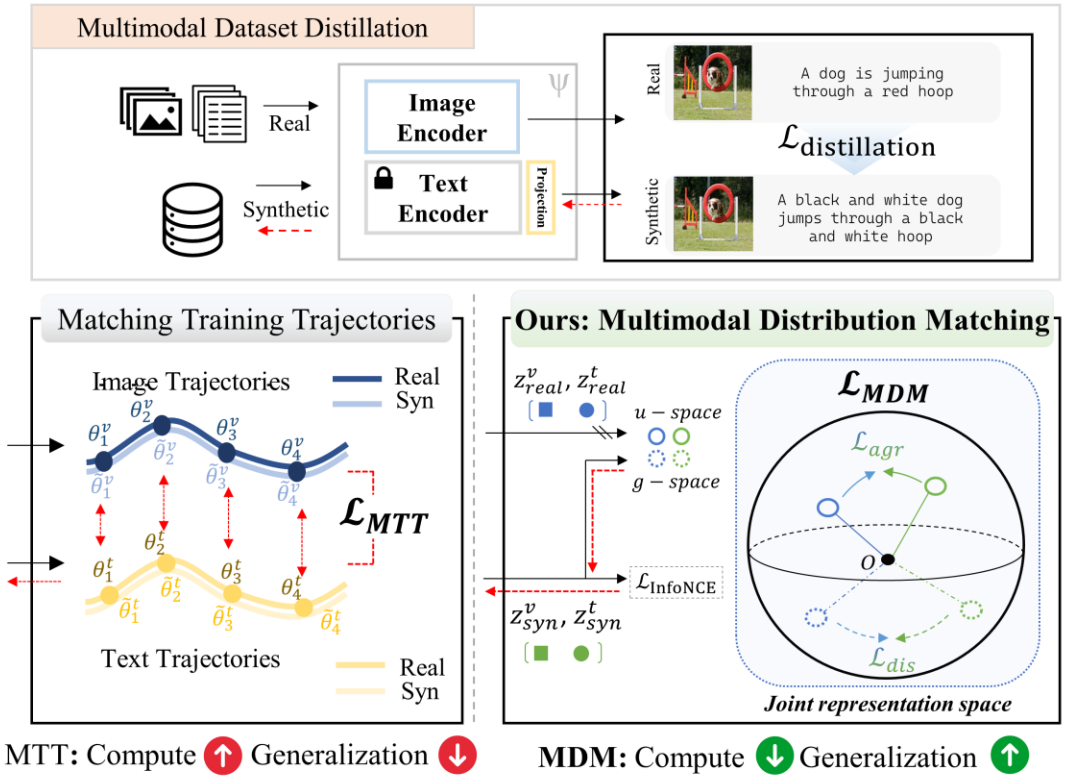


Dataset distillation



Matching Training Trajectory (MTT)

Motivation: Match Distributions, Not Trajectories

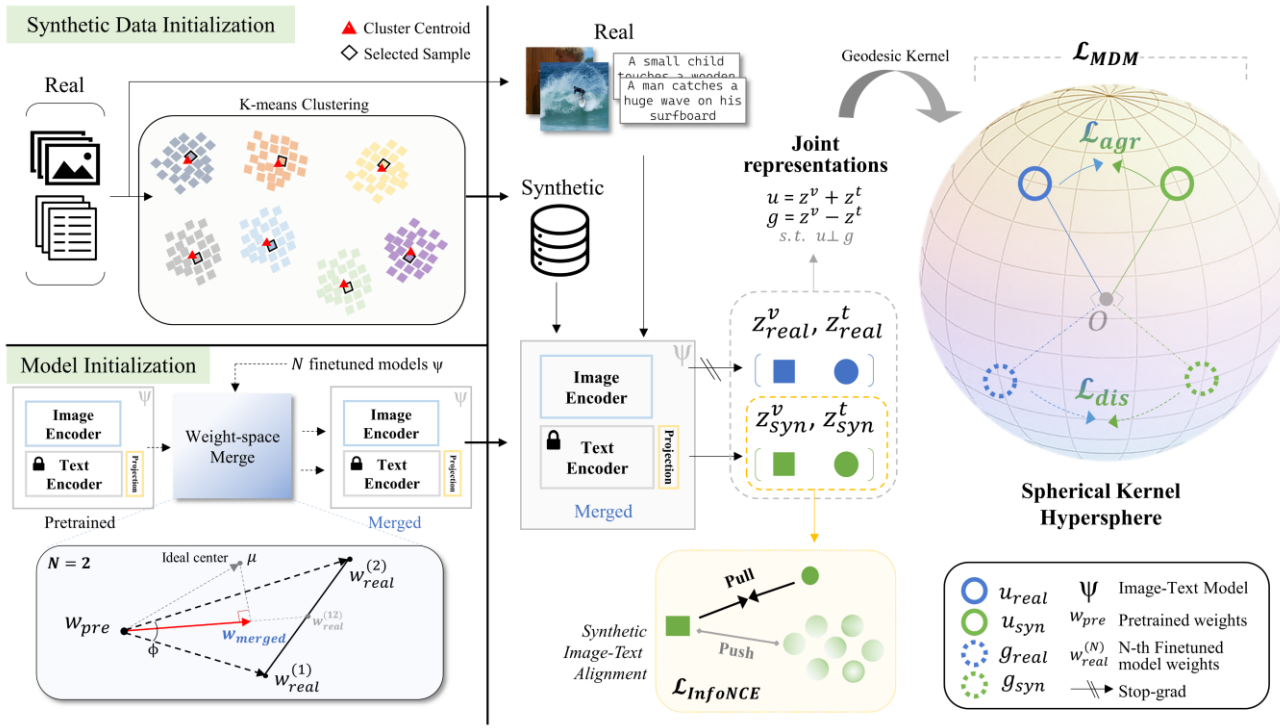


- ✓ **Desiderata for Effective Distillation**
  - ✓ **Directly align real and synthetic multimodal distributions**
    - Match synthetic pairs to the joint image-text representation distribution
    - Preserve both shared semantics and cross-modal structure
  - ✓ **Avoid Resource-heavy Trajectory Replay**
    - No need to reconstruct training dynamics or solve expensive bi-level optimization
    - Leads to substantially lower compute and memory cost
  - ✓ **Improve Architecture Robustness**
    - Distribution-level matching is less tied to a single source encoder geometry
    - Enables better cross-architecture generalization of distilled pairs

Multimodal Distribution Matching

✓ Three components of Multimodal Distribution Matching

1. **Synthetic Data Initialization**
   - Seed synthetic image-text pairs by K-means clustering in the joint embedding space, providing optimization a better starting point from representative multimodal prototypes
2. **Model Initialization**
   - Build a mixed image-text model by weight-space interpolation between a pretrained model and multiple finetuned experts, improving robustness to source-model bias
3. **Multimodal Distribution Matching**
   - Match real and synthetic pairs in the joint space using agreement and discrepancy features on the unit hypersphere, together with bidirectional contrastive loss for synthetic image-text alignment

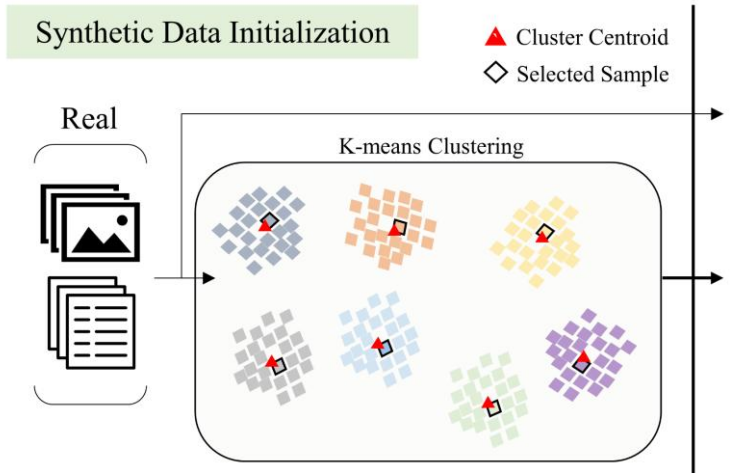


Overview of MDM.

Joint-space Synthetic Data Initialization

✓ Synthetic Data Initialization in the Joint Image-Text Space

- Embed each real pair using the concatenated image and text features after projection
- Run K-means clustering with $K = |D_{syn}|$ in the joint space, and initialize each synthetic pair with the sample nearest its cluster centroid in cosine distance
  → broader coverage of joint semantic modes while avoiding redundancy
  → stable and representative starting point
  → better reflects multimodal structure from the joint features



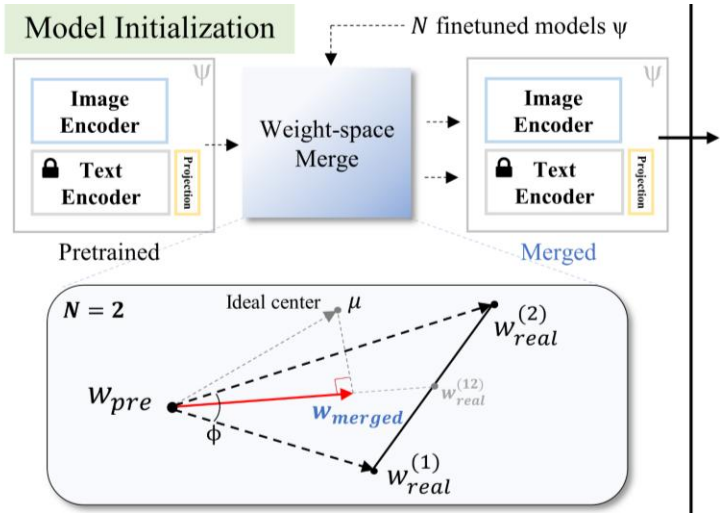
Synthetic Data Initialization in MDM

Model Initialization

✓ Model Initialization via Weight-space Interpolation

- Start from a pretrained image encoder and text projector as the anchor model
- Merge this anchor with multiple independently finetuned experts in weight space, rather than relying on a single finetuned model
- Use the agreement between expert displacement directions to decide how far to shift form the anchor:
  - More agreement → move toward experts
  - More disagreement → stay closer to the pretrained model

→ broader, less architecture-biased joint embedding space,
thus stabilizing distillation and improving cross-architecture generalization



Model Initialization in MDM

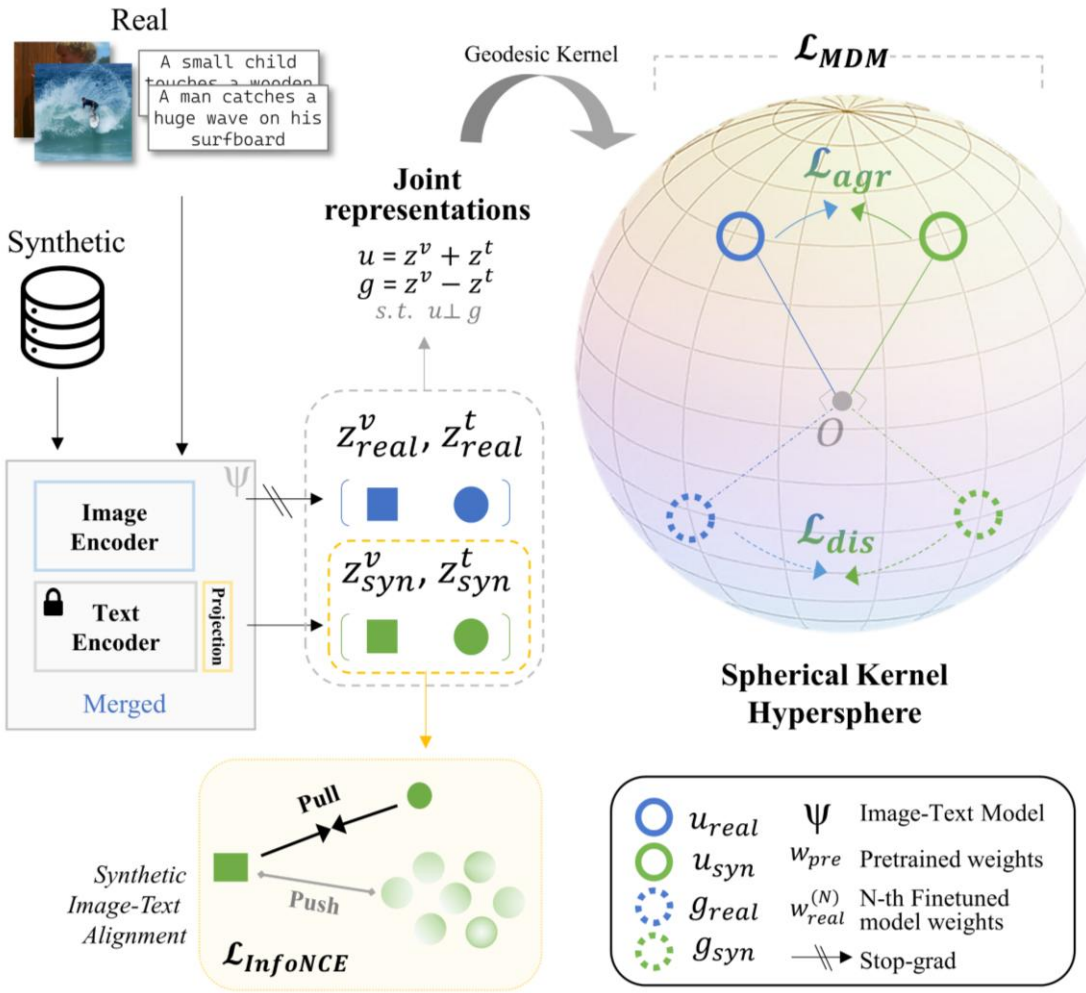
Geometry-Aware Multimodal Matching

✓ Geometry-Aware Multimodal Matching

- Construct two joint features from image and text embeddings:

  agreement $u = z^v + z^t$ and discrepancy $g = z^v - z^t$

- u captures modality-shared semantics such as objects, actions, and coarse scene layout, while g captures the modality gap between image and text
- Match the real and synthetic distributions of u and g on the unit hypersphere using geodesic kernel energy, which better respects spherical geometry than the Euclidean space
- Add bidirectional contrastive loss for synthetic image-text alignment



Multimodal Distribution Matching

Experimental Setup

✓ **Datasets**
- Flickr8k
- Flickr30k
- MS COCO

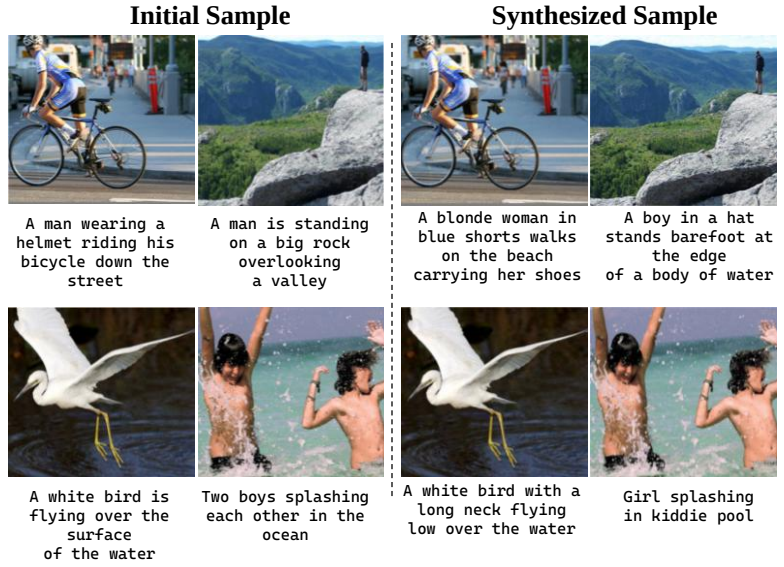
✓ **Task**
- Image-to-Text Retrieval at $K \in \{1,5,10\}$
- Text-to-Image Retrieval at $K \in \{1,5,10\}$

✓ **Synthetic Budget**
- N $\in \{100, 200,500\}$
- Each pair contains $3 \times 224 \times 224$ synthetic image and 768-dimensional synthetic text embedding

✓ **Baselines and Architectures**
- Coreset methods (Random, herding, K-center, forgetting)
- MDD baselines (MTT-VL, $\text{TESLA}_{\text{WBCE}}$, LoRS)
- Architectures: NFNet (Image), BERT (Text) with lightweight projection heads

Results: Image-Text Retrieval

- **Consistent gains over previous trajectory-based MDD**
  - Across Flickr8k, Flickr30k, and COCO and all budgets (100 / 200 / 500 pairs), MDM consistently outperforms MTT-VL and TESLA on bidirectional retrieval.
  - The gains are especially meaningful under extreme compression, showing that direct multimodal distribution matching remains effective even with very few synthetic pairs.
- **Competitive Performance Across the Board**
  - MDM is competitive with LoRS across the board, and notably surpasses LoRS on COCO at all budgets
  - On Flickr8k, MDM also achieves the best mean recall at 500 pairs



# Pairs	Dataset	Flickr8k							Flickr30k							COCO						
	Method	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	Mean	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	Mean	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	Mean
100 Pairs	MTT-VL	0.8	4.0	7.0	1.5	6.4	10.8	5.1	4.7	15.7	24.6	9.9	28.3	39.1	20.4	1.3	5.4	9.5	2.5	10.0	15.7	7.4
	$TESLA_{WBCE}$	1.2	4.7	7.9	4.8	14.7	22.9	9.4	0.5	2.3	4.7	5.5	19.5	28.9	10.2	0.3	1.0	1.8	2.0	7.7	13.5	4.4
	LoRS	4.9	18.0	29.0	7.0	22.8	34.8	19.4	8.3	24.1	35.1	11.8	35.8	49.2	**27.4**	1.8	7.1	12.2	3.3	12.2	19.6	9.4
	Ours	6.0	20.8	32.4	7.9	26.5	38.1	**21.9**	8.1	24.7	36.2	11.5	32.6	45.0	26.4	1.9	7.6	13.2	3.6	13.7	21.6	**10.3**
200 Pairs	MTT-VL	1.8	7.0	12.2	2.8	10.3	17.3	8.6	4.6	16.0	25.5	10.2	28.7	41.9	21.2	1.7	6.5	12.3	3.3	11.9	19.4	9.2
	$TESLA_{WBCE}$	1.2	4.7	8.4	6.6	19.5	29.5	11.7	0.2	1.3	2.5	2.8	10.4	17.4	5.8	0.1	0.2	0.5	0.7	3.1	5.3	1.7
	LoRS	6.3	20.5	31.6	9.5	26.3	38.2	22.1	8.6	25.3	36.6	14.5	38.7	53.4	**29.5**	2.4	9.3	15.5	4.3	14.2	22.6	11.4
	Ours	7.1	23.2	35.1	9.9	29.0	41.6	**24.3**	9.1	26.7	39.1	13.0	33.7	47.4	28.2	2.9	11.1	18.4	4.9	16.2	25.3	**13.1**
500 Pairs	MTT-VL	3.8	13.3	21.2	5.7	18.6	27.7	15.1	6.6	20.2	30.0	13.3	32.8	46.8	25.0	2.5	8.9	15.8	5.0	17.2	26.0	12.6
	$TESLA_{WBCE}$	2.5	8.8	14.1	6.9	19.6	29.0	13.5	1.1	7.3	12.6	5.1	15.3	23.8	10.9	0.8	3.6	6.7	1.7	5.9	10.2	4.8
	LoRS	6.9	22.0	33.1	10.9	31.0	45.8	25.0	10.0	28.9	41.6	15.5	29.8	53.7	**31.6**	2.8	9.9	16.5	5.3	18.3	27.9	13.5
	Ours	7.4	25.0	37.1	11.2	32.4	44.2	**26.2**	10.0	29.3	42.0	13.7	37.0	51.5	30.6	3.7	13.6	22.2	5.6	18.4	28.2	**15.3**

Results: Cross-Architecture Generalization

- ✓ **Generalization beyond the source encoder**
  - All synthetic datasets are distilled with NFNet+BERT, then evaluated on unseen image-text networks
  - Across all datasets and all budgets, MDM achieves higher cross-architecture performance than the MTT-based baseline, demonstrating stronger robustness to architecture changes

- ✓ **Why significant?**
  - Trajectory-matching baselines inherit source-architecture bias
  - MDM instead distills the data-level multimodal distribution, so the synthetic pairs transfer more reliably across models

# Pairs		Flickr8k							Flickr30k							COCO						
	Text	BERT			DistilBERT				BERT			DistilBERT				BERT			DistilBERT			
	Image	(a)	(b)	(c)	(a)	(b)	(c)	Mean	(a)	(b)	(c)	(a)	(b)	(c)	Mean	(a)	(b)	(c)	(a)	(b)	(c)	Mean
100	LoRS	19.4*	10.0	9.2	15.6	8.7	8.2	10.3	27.4*	6.5	7.1	24.0	5.8	6.1	9.9	9.4*	1.8	1.6	6.8	1.2	1.2	2.5
	Ours	21.9*	13.6	15.3	17.3	11.0	12.4	**13.9**	26.4*	13.7	18.9	20.8	11.6	15.3	**16.1**	10.3*	6.4	7.2	7.0	4.6	5.2	**6.1**
200	LoRS	22.1*	11.7	10.8	18.3	8.5	8.8	11.6	29.5*	10.0	10.9	22.7	8.3	8.9	12.2	11.4*	1.6	3.2	7.7	1.0	2.1	3.1
	Ours	24.3*	14.7	18.9	19.4	10.8	14.2	**15.6**	28.2*	15.0	20.8	23.3	11.7	16.3	**17.4**	13.1*	9.2	10.2	9.9	6.9	7.7	**8.7**
500	LoRS	25.0*	9.9	9.5	19.3	6.3	6.2	10.2	31.6*	15.3	13.4	22.1	10.8	10.7	14.5	13.5*	1.4	1.3	7.5	0.8	0.9	2.4
	Ours	26.2*	13.8	20.5	20.5	10.0	16.2	**16.2**	30.6*	17.4	23.3	23.9	13.5	18.5	**19.3**	15.3*	9.8	11.3	11.4	7.6	8.9	**9.8**

*source model (a)NFNet (b)NF-ResNet (c)NF-RegNet

Results: Compute Efficiency

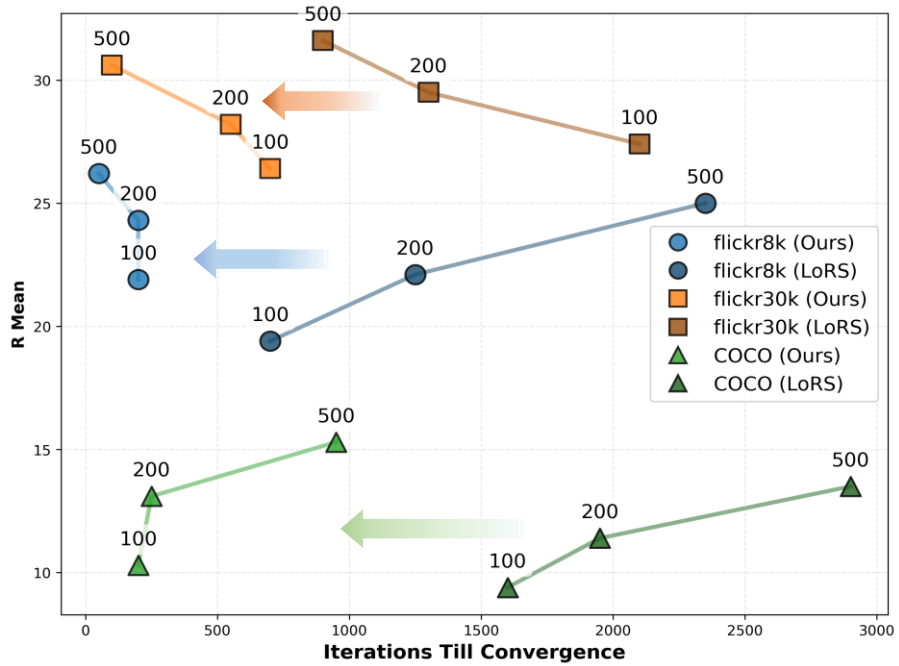
✓ **Significantly lower distillation cost than trajectory matching**

- Compared with LoRS, MDM reduces per-iteration distillation time from about 5.4 s to 1.7 s at 100 pairs, nearly a 3× speedup
- This comes from replacing bi-level trajectory replay with single-level distribution matching, avoiding repeated student trajectory reconstruction

✓ **Fewer iterations to strong performance**

- On Flickr8k, MDM reaches its best performance in 200 / 200 / 50 iterations for 100 / 200 / 500 pairs, versus 850 / 1250 / 2350 for LoRS
- This reduces total distillation time from 76.9 → 5.7 min, 137.5 → 8.0 min, and 206.4 → 3.7 min, corresponding to 93% / 94% / 98% savings

Flickr8k		Distillation		
Method	# Pairs	Time/iter (sec)	# iter till best	Total (min)
LoRS	100	5.43	850	76.93
Ours		1.72	200	5.73 (↓ 93%)
LoRS	200	6.60	1,250	137.50
Ours		2.39	200	7.97 (↓ 94%)
LoRS	500	5.27	2,350	206.41
Ours		4.41	50	3.68 (↓ 98%)



Empirical Findings

- ✓ **Component ablations**
  - Joint-space K-means data initialization yields the highest retrieval score over noise, random, image-only, text-only seeds
  - Weight-space interpolation outperforms pretrained-only, fine-tuned-only, and weighted-sum alternatives

- ✓ **Discrepancy modeling is especially important**
  - While agreement-discrepancy are complementary, discrepancy contributes more strongly than agreement

Data Init.

			K-means Clustering		
Syn. Data Init.	Noise	Random	Image	Text	Ours
IR	0.6	18.6	18.9	18.6	**19.7**
TR	0.5	22.6	22.7	22.8	**24.2**
Mean	0.5	20.6	20.8	20.7	**21.9**

Model Init.

Model Init.	Pre-trained	Fine-tuned	Weighted Sum	Ours
IR	5.1	14.3	17.6	**19.7**
TR	8.4	21.3	21.7	**24.2**
Mean	6.8	17.8	19.6	**21.9**

Multimodal Distribution Matching

No.	$\mathcal{L}_{\text{InfoNCE}}$	$\mathcal{L}_{\text{agr}}$	$\mathcal{L}_{\text{dis}}$	IR	TR	Mean
1	✓	✗	✗	18.81	23.15	20.98
2	✓	✓	✗	18.82	23.23	21.02
3	✓	✗	✓	19.22	23.84	21.53
4	✓	✓	✓	**19.73**	**24.15**	**21.94**

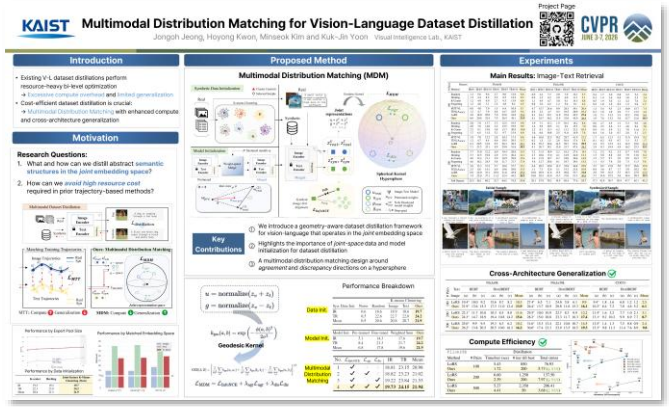
Key Takeaway

- ✓ **Match Distributions, Not Costly Trajectories**
  - MDM replaces expensive trajectory replay with direct multimodal distribution matching in the joint image-text space

- ✓ **Empirically Better Design Choices**
  - Joint-space K-Means gives representative multimodal prototypes for data initialization
  - Weight-space interpolation reduces source-model bias
  - Agreement and discrepancy matching together yields the strongest retrieval performance

- ✓ **Better transfer, lower distillation cost**
  - MDM achieves strong image-text retrieval, better cross-architecture generalization, and dramatically lower distillation cost than trajectory-based baselines

Multimodal Distribution Matching for Vision-Language Dataset Distillation



Project Page

Thank you!

Poster Session 4 (#153):
Thu, Jun 4, 2026 | 4:45 PM – 6:45 PM | ExHall A