Novelty Assessment Report

Paper: Asynchronous Matching with Dynamic Sampling for Multimodal Dataset Distillation
PDF URL: https://openreview.net/pdf?id=7SgSMKM2KF
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-07

Abstract

Multimodal Dataset Distillation (MDD) has emerged as a vital paradigm for enabling efficient training of vision-language models (VLMs) in the era of multimodal data proliferation. Unlike traditional dataset distillation methods that focus on single-modal tasks, MDD presents distinct challenges: (i) the effective distillation of heterogeneous multimodal knowledge, complicated by feature space misalignment and asynchronous optimization dynamics; and (ii) the lack of discrete class guidance, which hinders the distribution coverage and representativeness of synthetic data due to the vastness and continuity of the semantic space. To address these challenges, this paper proposes an Asynchronous Matching with Dynamic sampling (AMD) framework. AMD enables asynchronous trajectory matching by decoupling the selection of starting points for image and text trajectories. Additionally, a Semantics-Aware Prototype Mining module is introduced, which replaces random initialization by leveraging feature-space clustering to identify representative prototypes, enhancing the coverage and representativeness of the distilled samples. Extensive experiments demonstrate that AMD achieves superior distillation performance on Flickr30k and COCO (e.g., IR@1, IR@5, and IR@10 \textbf{gains of 4.5\%, 9.6\%, and 10.9\%}, respectively, on Flickr30k 200 pairs.) with negligible computational overhead.

Core Task Landscape

This paper addresses: **multimodal dataset distillation for vision-language models**

A total of **50 papers** were analyzed and organized into a taxonomy with **17 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Core Dataset Distillation Methods for Vision-Language Data**
- **Model Distillation and Compression for Vision-Language Models**
- **Cross-Modal Knowledge Transfer and Fusion**
- **Task-Specific Applications and Extensions**

Complete Taxonomy Tree

- multimodal dataset distillation for vision-language models Survey Taxonomy
- Core Dataset Distillation Methods for Vision-Language Data
  - Trajectory Matching and Gradient-Based Distillation ★ (3 papers)
  - [0] Asynchronous Matching with Dynamic Sampling for Multimodal Dataset Distillation (Anon et al., 2026) View paper
  - [1] Efficient multimodal dataset distillation via generative models (Zhao, 2025) View paper
  - [5] Multimodal Dataset Distillation for Image-Text Retrieval (Wu, 2023) View paper
  - Distribution and Prototype-Based Methods (2 papers)
  - [11] Dataset Distillation via Vision-Language Category Prototype (Zou Ya-wen, 2025) View paper
  - [36] Beyond Modality Collapse: Representations Blending for Multimodal Dataset Distillation (Zhang Xin, 2025) View paper
  - Generative and Latent-Space Distillation (1 papers)
  - [46] Latent Video Dataset Distillation (Li Ning, 2025) View paper
  - Scalability and Efficiency Advances (1 papers)
  - [32] The evolution of dataset distillation: Toward scalable and generalizable solutions (Liu Ping, 2025) View paper
- Model Distillation and Compression for Vision-Language Models
  - Multimodal Large Language Model Distillation (9 papers)
  - [2] Silkie: Preference distillation for large visual language models (Li Lei, 2023) View paper
  - [9] Llava-kd: A framework of distilling multimodal large language models (Cai Yu-Xuan, 2025) View paper
  - [12] Unlock the power: Competitive distillation for multi-modal large language models (Li XinWei, 2023) View paper
  - [21] Self-improving teacher cultivates better student: Distillation calibration for multimodal large language models (Xinwei Li, 2024) View paper
  - [24] Elevating Visual Perception in Multimodal LLMs with Visual Embedding Distillation (Jain, 2025) View paper
  - [28] CompoDistill: Attention Distillation for Compositional Reasoning in Multimodal LLMs (Kim, 2025) View paper
  - [29] Distilling multi-modal large language models for autonomous driving (Deepti Hegde, 2025) View paper
  - [35] Llavadi: What matters for multimodal large language models distillation (Xu, 2024) View paper
  - [37] MASSV: Multimodal Adaptation and Self-Data Distillation for Speculative Decoding of Vision-Language Models (Mahesan Ganesan, 2025) View paper
  - Vision-Language Encoder Distillation (7 papers)
  - [3] AMMKD: Adaptive multimodal multi-teacher distillation for lightweight vision-language models (Li YuQi, 2025) View paper
  - [8] DC-CLIP: Multilingual CLIP Compression via vision-language distillation and vision-language alignment (Wenbo Zhang, 2025) View paper

- [13] Layerwised multimodal knowledge distillation for vision-language pretrained model. (Jin Wang, 2024) View paper
- [17] Multimodal Adaptive Distillation for Leveraging Unimodal Encoders for Vision-Language Tasks (Wang, 2022) View paper
- [33] DIME-FM : DIstilling Multimodal and Efficient Foundation Models (Xi-Meng Sun, 2023) View paper
- [43] Distilled Dual-Encoder Model for Vision-Language Understanding (Liu Ming, 2022) View paper
- [49] Module-wise adaptive distillation for multimodality foundation models (Liang Chen, 2023) View paper
- Prompt and Adaptation Methods (3 papers)
- [4] PromptKD: Unsupervised Prompt Distillation for Vision-Language Models (Zheng Li, 2024) View paper
- [7] Preventing zero-shot transfer degradation in continual learning of vision-language models (Zangwei Zheng, 2023) View paper
- [41] Is less more? exploring token condensation as training-free test-time adaptation (Wang Zi-xin, 2025) View paper
- Early Exit and Efficiency Mechanisms (1 papers)
- [6] Multimodality Self-distillation for Fast Inference of Vision and Language Pretrained Models (Jun Kong, 2024) View paper

- Cross-Modal Knowledge Transfer and Fusion
  - Vision-Language Knowledge Distillation for Downstream Tasks (6 papers)
  - [10] Partdistill: 3d shape part segmentation by vision-language model distillation (Ardian Umam, 2024) View paper
  - [14] Enabling multimodal generation on clip via vision-language knowledge distillation (Dai, 2022) View paper
  - [16] Vldadaptor: Domain adaptive object detection with vision-language model distillation (Junjie Ke, 2024) View paper
  - [18] Filtering, Distillation, and Hard Negatives for Vision-Language Pre-Training (Filip RadenoviÄ, 2023) View paper
  - [22] Open-vocabulary Object Detection via Vision and Language Knowledge Distillation (Gu, 2021) View paper
  - [44] KDNet: Leveraging Vision-Language Knowledge Distillation for Few-Shot Object Detection (Mengyuan Ma, 2024) View paper
  - Multimodal Fusion and Alignment (3 papers)
  - [34] Align before Fuse: Vision and Language Representation Learning with Momentum Distillation (Junnan Li, 2022) View paper
  - [39] TransferCVLM: Transferring Cross-Modal Knowledge for Vision-Language Modeling (Choi, 2024) View paper
  - [40] Knowledge Distillation Across Vision and Language (Zhiyuan Fang, 2023) View paper
  - Cross-Lingual and Cross-Domain Adaptation (1 papers)
  - [25] Cross-Lingual Adaptation for Vision-Language Model via Multimodal Semantic Distillation (Yu Weng, 2025) View paper
- Task-Specific Applications and Extensions
  - Visual Question Answering and Reasoning (3 papers)
  - [20] Improving context understanding in multimodal large language models via multimodal composition learning (W Li, 2024) View paper
  - [31] Multimodal commonsense knowledge distillation for visual question answering (student abstract) (Han, 2025) View paper
  - [38] Knowledge condensation and reasoning for knowledge-based VQA (Jia Jian, 2024) View paper
  - Multimodal Retrieval and Hashing (2 papers)
  - [26] Unsupervised graph reasoning distillation hashing for multimodal hamming space search with vision-language model (Lina Sun, 2024) View paper
  - [47] Unveil: Unified Visual-Textual Integration and Distillation for Multi-modal Document Retrieval (Guo Jiayan, 2025) View paper
  - Action Recognition and Video Understanding (2 papers)
  - [42] Vision-language meets the skeleton: Progressively distillation with cross-modal knowledge for 3d action representation learning (Yang Chen, 2024) View paper
  - [45] Gpt4video: A unified multimodal large language model for lnstruction-followed understanding and safety-aware generation (Zhanyu Wang, 2024) View paper
  - Emotion Recognition and Sentiment Analysis (1 papers)
  - [23] Decoupled Multimodal Distilling for Emotion Recognition (Yong Li, 2023) View paper
  - Misinformation Detection and Safety (2 papers)
  - [19] Distilling implicit multimodal knowledge into large language models for zero-resource dialogue generation (Bo Zhang, 2025) View paper
  - [50] Self-supervised distilled learning for multi-modal misinformation identification (Michael Mu, 2023) View paper
  - Specialized Applications (4 papers)
  - [15] Visual Program Distillation: Distilling Tools and Programmatic Reasoning into Vision-Language Models (Yushi Hu, 2024) View paper
  - [27] MMT-ARD: Multimodal Multi-Teacher Adversarial Distillation for Robust Vision-Language Models (Yuqi Li, 2025) View paper
  - [30] KAID: Knowledge-Aware Interactive Distillation for Vision-Language Models (Da Zhang, 2025) View paper
  - [48] Contextual Paralinguistic Data Creation for Multi-Modal Speech-LLM: Data Condensation and Spoken QA Generation (Wang Qiongqiong, 2025) View paper

Narrative

Core task: multimodal dataset distillation for vision-language models. The field organizes around four main branches. Core Dataset Distillation Methods for Vision-Language Data focuses on condensing large-scale multimodal datasets into compact, representative subsets using trajectory matching, gradient-based techniques, and other distillation strategies—works like Multimodal Dataset Distillation[5] and Efficient Multimodal Distillation[1] exemplify this direction. Model Distillation and Compression for Vision-Language Models emphasizes reducing model size and computational cost by transferring knowledge from large teacher models to smaller students, often leveraging cross-modal alignment and layer-wise distillation (e.g., AMMKD[3], PromptKD[4]). Cross-Modal Knowledge Transfer and Fusion explores how to effectively align and fuse information across vision and language modalities, addressing challenges like modality imbalance and zero-shot transfer (e.g., Preventing Zero-shot Degradation[7], Align before Fuse[34]). Task-Specific Applications and Extensions applies distillation techniques to downstream problems such as open-vocabulary detection, autonomous driving, and video understanding, demonstrating the practical utility of these methods.

A particularly active line of work within Core Dataset Distillation Methods centers on trajectory matching and gradient-based distillation, where the goal is to synthesize small datasets that mimic the training dynamics of full-scale data. Asynchronous Matching Multimodal[0] sits squarely in this cluster, addressing the challenge of aligning asynchronous gradient trajectories across vision and language modalities—a key bottleneck when distilling multimodal data. Compared to Multimodal Dataset Distillation[5], which introduced foundational techniques for multimodal condensation, Asynchronous Matching Multimodal[0] emphasizes temporal alignment and modality-specific optimization schedules. Meanwhile, Efficient Multimodal Distillation[1] explores computational efficiency in similar settings, highlighting trade-offs between distillation quality and resource constraints. These works collectively push toward scalable, high-fidelity dataset distillation, though open questions remain about generalization across diverse vision-language architectures and task distributions.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Efficient multimodal dataset distillation via generative models

Authors: Zhao, Zhenghao, Wang, Haoxuan, Zhenghao Zhao, et al. (14 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Dataset distillation aims to synthesize a small dataset from a large dataset, enabling the model trained on it to perform well on the original dataset. With the blooming of large language models and multimodal large language models, the importance of multimodal datasets, particularly image-text datasets, has grown significantly. However, existing multimodal dataset distillation methods are constrained by the Matching Training Trajectories algorithm, which significantly increases the computing re...

Relationship Analysis

Both papers belong to the Trajectory Matching and Gradient-Based Distillation category, addressing multimodal dataset distillation for vision-language models through trajectory-based approaches. They overlap in their focus on matching expert trajectories for image-text pairs and addressing the challenges of cross-modal alignment in dataset distillation. However, the original paper (AMD) proposes asynchronous trajectory matching that decouples image and text trajectory sampling with dynamic sampling ranges, while the candidate paper (EDGE) shifts away from trajectory matching entirely, instead leveraging generative diffusion models with contrastive and diversity losses to achieve faster distillation with lower computational overhead.

---

2. Multimodal Dataset Distillation for Image-Text Retrieval

Authors: Wu, Xindi, Xindi Wu, Zhiwei Deng, Deng Zhiwei, et al. (8 authors total) | **Year/Venue**: 2023 | **URL**: View paper

Abstract

Dataset distillation methods reduce large-scale datasets to smaller sets of synthetic data, preserving sufficient information to quickly train a new model from scratch. However, prior work on dataset distillation has focused exclusively on image classification datasets, whereas modern large-scale datasets are primarily vision-language datasets. In this work, we design the first vision-language dataset distillation method, building on the idea of trajectory matching. A key challenge is that visio...

Relationship Analysis

Both papers belong to the Trajectory Matching and Gradient-Based Distillation category, focusing on matching training trajectories between models trained on original and synthetic datasets for multimodal dataset distillation. They overlap in applying trajectory matching to vision-language models using contrastive learning objectives and addressing the challenge of distilling image-text pairs without discrete class labels. The key difference is that the original paper (AMD) introduces asynchronous trajectory matching that decouples image and text trajectory sampling points and proposes semantics-aware prototype mining for initialization, while the candidate paper (MTT-VL) uses synchronous bi-trajectory matching where image and text trajectories are sampled from the same training steps and employs random initialization from real samples.

Contributions Analysis

Overall novelty summary. The paper proposes an Asynchronous Matching with Dynamic sampling (AMD) framework for multimodal dataset distillation, targeting vision-language models. It resides in the 'Trajectory Matching and Gradient-Based Distillation' leaf, which contains only three papers total, including this work and two siblings. This indicates a relatively sparse research direction within the broader taxonomy of 50 papers across 36 topics. The focus on asynchronous trajectory matching and semantics-aware prototype mining positions the work at the intersection of trajectory-based distillation and multimodal optimization challenges.

The taxonomy tree reveals that the paper's immediate neighbors address foundational multimodal distillation techniques and efficiency concerns, while adjacent leaves explore distribution-based methods, generative approaches, and scalability advances. The broader 'Core Dataset Distillation Methods' branch sits alongside three other major directions: model compression for VLMs, cross-modal knowledge transfer, and task-specific applications. The paper's emphasis on asynchronous optimization and prototype mining distinguishes it from distribution-matching methods in neighboring leaves, though both address the challenge of synthesizing representative multimodal data without discrete class labels.

Among 18 candidates examined across three contributions, no clearly refutable prior work was identified. The Asynchronous Matching framework examined 4 candidates with 0 refutations, Semantics-Aware Prototype Mining examined 8 candidates with 0 refutations, and the MMD-based dynamic sampling strategy examined 6 candidates with 0 refutations. This suggests that within the limited search scope—top-K semantic matches plus citation expansion—the specific combination of asynchronous trajectory decoupling and feature-space clustering for prototype initialization appears not to have direct precedents. However, the analysis explicitly notes this is not an exhaustive literature search.

Based on the limited examination of 18 candidates, the work appears to introduce novel mechanisms for handling multimodal distillation challenges, particularly the asynchronous optimization dynamics and prototype-based initialization. The sparse population of its taxonomy leaf and absence of refutable candidates within the search scope suggest potential novelty, though the small scale of the literature search means substantial related work may exist beyond the examined set. The contribution's distinctiveness hinges on the specific integration of asynchronous matching with semantics-aware mining rather than individual components.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Asynchronous Matching with Dynamic Sampling (AMD) Framework

Description: The authors propose a novel framework that decouples the sampling of image and text expert trajectories during multimodal dataset distillation. This asynchronous matching strategy addresses the inherent heterogeneity in learning dynamics between visual and text modalities, allowing more flexible combinations of parameters from different training epochs to improve synthetic data optimization.

This contribution was assessed against **4 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Transformer-GAN hybrid architecture for cross-modal virtual-real alignment in intelligent manufacturing system design

URL: View paper

Brief Assessment

Transformer-GAN Hybrid[67] focuses on cross-modal virtual-real alignment in manufacturing systems (vision, RF, vibration data), not multimodal dataset distillation for vision-language models. The technical domains are fundamentally different.

---

2. VIDEO GENERATION AND UNDERSTANDING WITH MULTIMODAL LEARNING

URL: View paper

Brief Assessment

Video Generation Understanding[70] focuses on video generation and understanding tasks with motion/mixed trajectory distillation, not multimodal dataset distillation for image-text pairs with asynchronous trajectory matching.

---

3. Dataset Distillation in the Era of Large-Scale Data: Methods, Analysis, and Future Directions

URL: View paper

Brief Assessment

Dataset Distillation Era[69] is a survey paper covering dataset distillation methods broadly. The provided context fragments do not contain specific technical details about asynchronous trajectory matching or decoupled sampling strategies for multimodal distillation that would refute the original paper's novelty claim.

---

4. From Models to Systems: A Comprehensive Survey of Efficient Multimodal Learning

URL: View paper

Brief Assessment

Efficient Multimodal Learning[68] focuses on general multimodal learning efficiency techniques including knowledge distillation, not specifically on multimodal dataset distillation with trajectory matching for synthetic data generation.

Contribution 2: Semantics-Aware Prototype Mining (SPM) Module

Description: The authors introduce a module that performs clustering in the joint semantic feature space to identify representative sample prototypes. These prototypes replace randomly selected initial points and are used to initialize the synthesis process, substantially enhancing the diversity and representativeness of distilled samples without discrete class guidance.

This contribution was assessed against **8 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Mine-distill-prototypes for complete few-shot class-incremental learning in image classification

URL: View paper

Brief Assessment

Mine-distill-prototypes[58] focuses on few-shot class-incremental learning with feature distillation to eliminate ineffective features, not on clustering-based prototype mining for initializing dataset distillation synthesis in multimodal contexts.

---

2. Label-Guided relation prototype generation for Continual Relation Extraction

URL: View paper

Brief Assessment

Label-Guided Relation Prototype[53] focuses on continual relation extraction with label-guided prototype generation for relation classification tasks, while the original paper addresses multimodal dataset distillation through clustering-based prototype mining in joint image-text feature spaces. These are fundamentally different problem domains with distinct technical approaches.

---

3. Multi-granularity class prototype topology distillation for class-incremental source-free unsupervised domain adaptation

URL: View paper

Brief Assessment

Multi-granularity Class Prototype[51] focuses on class-incremental domain adaptation with prototype mining for positive class identification in domain shift scenarios, not on initializing dataset distillation synthesis for multimodal data without discrete class guidance.

---

4. Feature Selection, Clustering, and Prototype Placement for Turbulence Datasets

URL: View paper

Brief Assessment

Prototype Placement Turbulence[59] focuses on turbulence flow data clustering and prototype selection for physics simulations, not multimodal dataset distillation or vision-language model training.

---

5. Prokd: an unsupervised prototypical knowledge distillation network for zero-resource cross-lingual named entity recognition

URL: View paper

Brief Assessment

Prokd[56] focuses on cross-lingual NER using prototypes for knowledge distillation between languages, not for initializing multimodal dataset distillation synthesis. The clustering serves a fundamentally different purpose in a different task domain.

---

6. Feature Distillation-Based Uniformity Few-Shot Domain Adaptation for Cross-Domain Fault Diagnosis With Sample Shortage

URL: View paper

Brief Assessment

Feature Distillation Uniformity[55] focuses on fault diagnosis in mechanical systems using prototypical networks with uniformity principles, not multimodal dataset distillation or clustering-based initialization for synthesis processes.

---

7. Pcps: Patient cardiac prototypes to probe ai-based medical diagnoses, distill datasets, and retrieve patients

URL: View paper

Brief Assessment

Pcps[57] focuses on learning patient-specific prototypes for medical diagnosis in cardiac data, not on clustering-based initialization for multimodal dataset distillation synthesis. The clustering approach in Pcps[57] serves to create patient representations for diagnosis, not to initialize synthetic data generation.

---

8. Diversified semantic distribution matching for dataset distillation

URL: View paper

Brief Assessment

Diversified Semantic Distribution[52] focuses on distribution matching via covariance matrices for single-modal classification tasks, not clustering-based prototype mining for multimodal dataset initialization.

Contribution 3: Maximum Mean Discrepancy Based Dynamic Sampling Strategy

Description: The authors develop a data-driven sampling strategy that uses Maximum Mean Discrepancy to quantify parameter update magnitudes between consecutive epochs. This approach adaptively establishes differential sampling ranges for visual and text modalities based on their relative convergence dynamics, preventing excessive asynchronicity while capturing inter-modal learning speed discrepancies.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Research on data augmentation algorithms for few-shot image classification based on generative adversarial networks

URL: View paper

Brief Assessment

Few-shot GAN Augmentation[62] focuses on data augmentation for image classification using GANs, not on multimodal trajectory matching or convergence dynamics between vision and language modalities.

2. Multimodal Cross-Domain Recommendation Guided by Optimal Transport with Adaptive Confidence Thresholding Estimator (ChinaMM)

URL: View paper

Brief Assessment

Multimodal Cross-Domain Recommendation[66] focuses on cross-domain recommendation systems with optimal transport, not multimodal dataset distillation with trajectory matching and convergence dynamics.

3. Sampling with Adaptive Variance for Multimodal Distributions

URL: View paper

Brief Assessment

Adaptive Variance Sampling[65] focuses on adaptive diffusion coefficients for sampling multimodal distributions in continuous optimization contexts, not on dataset distillation or cross-modal trajectory matching for vision-language models.

4. Multimodal generative learning utilizing jensen-shannon-divergence

URL: View paper

Brief Assessment

Multimodal Generative Learning[61] focuses on multimodal generative learning using Jensen-Shannon divergence for variational inference in VAEs, not on adaptive sampling strategies for dataset distillation or convergence dynamics quantification using MMD.

5. Continuous adaptive path sampling for efficient multimodal sampling and marginalization

URL: View paper

Brief Assessment

Continuous Adaptive Path[64] focuses on normalizing constant estimation and tempering for Bayesian inference, not multimodal dataset distillation with visual-text trajectory matching. The MMD usage contexts are fundamentally different.

6. Mental sampling in multimodal representations

URL: View paper

Brief Assessment

Mental Sampling Multimodal[63] focuses on sampling algorithms (MC^3) for mental representations in cognitive tasks, not on multimodal dataset distillation or vision-language model training. The MMD usage in Mental Sampling Multimodal[63] is not discussed in the provided context.

Appendix: Text Similarity Detection

Textual similarity detection checked 22 papers and found 2 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Multimodal Dataset Distillation for Image-Text Retrieval

Detected in: Core Task (sibling)

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Asynchronous Matching with Dynamic Sampling for Multimodal Dataset Distillation View paper
- [1] Efficient multimodal dataset distillation via generative models View paper
- [2] Silkie: Preference distillation for large visual language models View paper
- [3] AMMKD: Adaptive multimodal multi-teacher distillation for lightweight vision-language models View paper
- [4] PromptKD: Unsupervised Prompt Distillation for Vision-Language Models View paper
- [5] Multimodal Dataset Distillation for Image-Text Retrieval View paper
- [6] Multimodality Self-distillation for Fast Inference of Vision and Language Pretrained Models View paper
- [7] Preventing zero-shot transfer degradation in continual learning of vision-language models View paper
- [8] DC-CLIP: Multilingual CLIP Compression via vision-language distillation and vision-language alignment View paper
- [9] Llava-kd: A framework of distilling multimodal large language models View paper
- [10] Partdistill: 3d shape part segmentation by vision-language model distillation View paper
- [11] Dataset Distillation via Vision-Language Category Prototype View paper
- [12] Unlock the power: Competitive distillation for multi-modal large language models View paper
- [13] Layerwised multimodal knowledge distillation for vision-language pretrained model. View paper
- [14] Enabling multimodal generation on clip via vision-language knowledge distillation View paper
- [15] Visual Program Distillation: Distilling Tools and Programmatic Reasoning into Vision-Language Models View paper

- [16] Vldadaptor: Domain adaptive object detection with vision-language model distillation View paper
- [17] Multimodal Adaptive Distillation for Leveraging Unimodal Encoders for Vision-Language Tasks View paper
- [18] Filtering, Distillation, and Hard Negatives for Vision-Language Pre-Training View paper
- [19] Distilling implicit multimodal knowledge into large language models for zero-resource dialogue generation View paper
- [20] Improving context understanding in multimodal large language models via multimodal composition learning View paper
- [21] Self-improving teacher cultivates better student: Distillation calibration for multimodal large language models View paper
- [22] Open-vocabulary Object Detection via Vision and Language Knowledge Distillation View paper
- [23] Decoupled Multimodal Distilling for Emotion Recognition View paper
- [24] Elevating Visual Perception in Multimodal LLMs with Visual Embedding Distillation View paper
- [25] Cross-Lingual Adaptation for Vision-Language Model via Multimodal Semantic Distillation View paper
- [26] Unsupervised graph reasoning distillation hashing for multimodal hamming space search with vision-language model View paper
- [27] MMT-ARD: Multimodal Multi-Teacher Adversarial Distillation for Robust Vision-Language Models View paper
- [28] CompoDistill: Attention Distillation for Compositional Reasoning in Multimodal LLMs View paper
- [29] Distilling multi-modal large language models for autonomous driving View paper
- [30] KAID: Knowledge-Aware Interactive Distillation for Vision-Language Models View paper
- [31] Multimodal commonsense knowledge distillation for visual question answering (student abstract) View paper
- [32] The evolution of dataset distillation: Toward scalable and generalizable solutions View paper
- [33] DIME-FM : DIstilling Multimodal and Efficient Foundation Models View paper
- [34] Align before Fuse: Vision and Language Representation Learning with Momentum Distillation View paper
- [35] Llavadi: What matters for multimodal large language models distillation View paper
- [36] Beyond Modality Collapse: Representations Blending for Multimodal Dataset Distillation View paper
- [37] MASSV: Multimodal Adaptation and Self-Data Distillation for Speculative Decoding of Vision-Language Models View paper
- [38] Knowledge condensation and reasoning for knowledge-based VQA View paper
- [39] TransferCVLM: Transferring Cross-Modal Knowledge for Vision-Language Modeling View paper
- [40] Knowledge Distillation Across Vision and Language View paper
- [41] Is less more? exploring token condensation as training-free test-time adaptation View paper
- [42] Vision-language meets the skeleton: Progressively distillation with cross-modal knowledge for 3d action representation learning View paper
- [43] Distilled Dual-Encoder Model for Vision-Language Understanding View paper
- [44] KDNet: Leveraging Vision-Language Knowledge Distillation for Few-Shot Object Detection View paper
- [45] Gpt4video: A unified multimodal large language model for lnstruction-followed understanding and safety-aware generation View paper
- [46] Latent Video Dataset Distillation View paper
- [47] Unveil: Unified Visual-Textual Integration and Distillation for Multi-modal Document Retrieval View paper
- [48] Contextual Paralinguistic Data Creation for Multi-Modal Speech-LLM: Data Condensation and Spoken QA Generation View paper
- [49] Module-wise adaptive distillation for multimodality foundation models View paper
- [50] Self-supervised distilled learning for multi-modal misinformation identification View paper
- [51] Multi-granularity class prototype topology distillation for class-incremental source-free unsupervised domain adaptation View paper
- [52] Diversified semantic distribution matching for dataset distillation View paper
- [53] Label-Guided relation prototype generation for Continual Relation Extraction View paper
- [54] Prototype-Decomposed Knowledge Distillation for Learning Generalized Federated Representation View paper
- [55] Feature Distillation-Based Uniformity Few-Shot Domain Adaptation for Cross-Domain Fault Diagnosis With Sample Shortage View paper
- [56] Prokd: an unsupervised prototypical knowledge distillation network for zero-resource cross-lingual named entity recognition View paper
- [57] Pcps: Patient cardiac prototypes to probe ai-based medical diagnoses, distill datasets, and retrieve patients View paper
- [58] Mine-distill-prototypes for complete few-shot class-incremental learning in image classification View paper
- [59] Feature Selection, Clustering, and Prototype Placement for Turbulence Datasets View paper
- [60] Simple yet Effective Graph Distillation via Clustering View paper
- [61] Multimodal generative learning utilizing jensen-shannon-divergence View paper
- [62] Research on data augmentation algorithms for few-shot image classification based on generative adversarial networks View paper
- [63] Mental sampling in multimodal representations View paper
- [64] Continuous adaptive path sampling for efficient multimodal sampling and marginalization View paper
- [65] Sampling with Adaptive Variance for Multimodal Distributions View paper
- [66] Multimodal Cross-Domain Recommendation Guided by Optimal Transport with Adaptive Confidence Thresholding Estimator (ChinaMM) View paper
- [67] Transformer-GAN hybrid architecture for cross-modal virtual-real alignment in intelligent manufacturing system design View paper
- [68] From Models to Systems: A Comprehensive Survey of Efficient Multimodal Learning View paper
- [69] Dataset Distillation in the Era of Large-Scale Data: Methods, Analysis, and Future Directions View paper
- [70] VIDEO GENERATION AND UNDERSTANDING WITH MULTIMODAL LEARNING View paper