SCALABLE LIFELONG MULTIMODAL INSTRUCTION TUNING VIA DYNAMIC DATA SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual instruction datasets from various distributors are released at different times and often contain a significant number of redundant text-image pairs, depending on their task compositions (*i.e.*, skills) or reference sources. This redundancy greatly limits the efficient deployment of **lifelong adaptable** Multimodal Large Language Models (MLLMs), hindering their ability to refine existing skills and acquire new competencies over time. To address this, we reframe the problem of Lifelong Instruction Tuning (LiIT) via data selection, where the model automatically selects beneficial samples to learn from earlier and new datasets based on the current state of acquired knowledge in the model. Based on empirical analyses which show that selecting the best data subset using a static importance measure is often ineffective for multi-task datasets with evolving distributions, we propose LAMP, a new multi-way and adaptive data selection approach that dynamically balances sample efficiency and effectiveness during LiIT. We first construct pseudo-skill clusters by grouping gradient-based sample vectors. Next, we select the best-performing data selector for each skill cluster from a pool of selector experts, including our newly proposed scoring function, *Image Grounding score*. This data selector samples a subset of the most important samples from each skill cluster for training. To prevent the continuous increase in the size of the dataset pool during LiIT, which would result in excessive computation, we further introduce a cluster-wise permanent data pruning strategy to remove the most semantically redundant samples from each cluster, keeping computational requirements manageable. We validate the effectiveness and efficiency of LAMP over a sequence of various multimodal instruction tuning datasets with various tasks, including (Knowledge) VQA, multilingual, grounding, reasoning, language-only, and multi-image comprehension tasks. Training with samples selected by LAMP alleviates catastrophic forgetting, especially for rare tasks, and promotes forward transfer across the continuum using only a fraction of the original datasets.

1 INTRODUCTION

Multimodal instruction tuning (Liu et al., 2023b; Zhang et al., 2023a; Liu et al., 2023a; Gan et al., 2024; Yoon et al., 2024) has been actively explored to enhance visual reasoning or the generation ability of Multimodal Large Language Models (MLLMs) (Zhu et al., 2023; Li et al., 2023a; Tang et al., 2023; Team et al., 2023; Munasinghe et al., 2023; Yu et al., 2024; Zhang et al., 2024) following user intents by training models on human or machine-generated multi-task visual instruction tuning datasets (Li et al., 2023c; Yin et al., 2023; Xu et al., 2024; Chen et al., 2024). While many distributors continue to release new high-quality instruction-tuning tasks and datasets, continually adapting large models to these massive datasets over time is prohibitively costly and inefficient. Given a pre-trained MLLM and the *continuous expansion of the dataset pool* with a stream of instruction-tuning datasets, as commonly observed in the research community today, the challenge lies in developing an ever-evolving, instruction-following MLLM in **the most data- and computation-efficient manner**. This research question poses a realistic, sustainable instruction tuning scenario for MLLMs, distinct from conventional continual learning (Zenke et al., 2017; Yoon et al., 2018; Van de Ven & Tolias, 2019), which focuses on learning a sequence of disjoint tasks. Specifically, at each time step, we assume a new multimodal instruction-tuning dataset is added to the existing training pool, which already contains previous datasets. The model's goal is to train on samples from this continually expanding

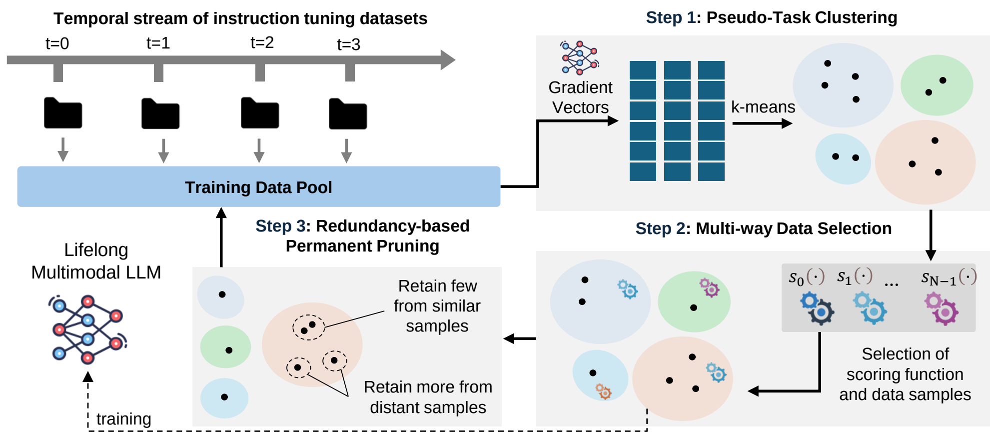

Figure 1: **Illustration of LAMP**. When a new dataset is incorporated into the data pool at the beginning of each timestep, LAMP extracts sample vectors and forms pseudo-task clusters based on their similarity. Using a set of scoring functions, LAMP predicts the most suitable scoring function for each cluster and train an MLLM on the selected samples. To prevent excessive computation as the pool size grows, we introduce dataset compression by permanently removing redundant samples.

dataset pool, a scenario we refer to as *Lifelong Instruction Tuning (LiIT)*. This scenario is critical as recent datasets tend to be exhaustively large to cover a variety of skills in uni- and multimodal tasks.

Our initial experiments with *sequential* multimodal instruction tuning, *i.e.*, training on a sequence of instruction tuning datasets, show catastrophic forgetting, especially when new output modalities like bounding boxes and key points are introduced. While experience replay on a small subset of past datasets alleviates forgetting to some extent, it is insufficient for retaining rare or unique tasks that appear only once, due to the skewed task distribution. This issue is compounded by the loosely defined nature of 'tasks' in instruction tuning (*e.g.*, multi-turn conversations across multiple tasks on the same image) and the lack of sample-wise task labels. Therefore, we address the problem of LiIT from a data selection perspective, enabling the model to learn skills in a balanced way over time and avoid overfitting to dominant tasks. Specifically, we explore how to prune past and incoming datasets to create a balanced training set at each time step, considering the model's current state.

To build an efficient lifelong-evolving MLLM, we introduce *LAMP: Lifelong and Adaptive Multi-way Pruning*, an efficient and dynamic multimodal data selection strategy. At each time step, LAMP selects the most beneficial samples for the current model from the data pool, adapting to the model's evolving knowledge and changing dataset distributions. This is crucial for LiIT, as sample importance shifts over time. LAMP operates in two major steps: (1) **Task-based clustering.** When a new dataset is introduced, we integrate it into the training data pool and create pseudo-task clusters using gradient vectors to represent data samples. (2) **Cluster-wise data selection.** We select influential samples from each cluster for model training. Motivated by the observation that different score functions (Paul et al., 2021; Toneva et al.; Coleman et al.) define sample importance differently, we propose a new **multi-way data selection approach** that chooses the best scoring function from a pool of experts based on its discriminativeness (measured by entropy). This selects a skill-balanced subset of highly influential samples from the current training data pool. As the size of the data pool continues to grow with time, the inference step for computing scores and data representations in LAMP, as in most existing data selection strategies, can become prohibitively expensive. Hence, to maintain a sufficiently diverse yet computationally manageable data pool, LAMP performs an additional step: (3) **Permanent data pruning** that removes semantically redundant samples from the data pool at the end of each time step, thereby continually controlling its size during LiIT.

To better assess the influence of multimodal samples, we also propose a new scoring function, called the *image grounding score* (IG), that measures the relative change in sample perplexity when the model is grounded by visual information. This metric prioritizes samples that effectively utilize and improve the multimodal skills of the MLLM and serves as an effective data selector in LAMP.

We design the experimental setup for this previously unexplored scenario of lifelong multimodal instruction tuning using the pre-trained LLaVA 1.5 (Liu et al., 2023a) model over a stream of five visual instruction tuning datasets. As discussed earlier, experience replay is insufficient for alleviating forgetting, however, a randomly selected subset from the data pool of past and new datasets brings down the forgetting rate from 26% to nearly 2%. Score-based data-selection strategies largely fail in this setting due to their inability to select task-balanced data subsets from multi-task datasets. LAMP not only minimizes the forgetting to a mere 0.9% using only a fraction of the training data pool, but also promotes forward transfer of skills and consistently achieves >100% relative gains.

LAMP provides an intuitive framework for dynamic data selection in the temporal scenario. We conduct extensive ablations of LAMP to find its best-performing settings. One of our significant findings is that hidden layer outputs represent the semantic component whereas gradient vectors represent the skill component of samples. Hence, gradient vectors are more effective at pooling samples into pseudo-task clusters (see examples in Figures 2 and 4). Further, we find that gradients from the middle layer of the model lead to best overall performance, suggesting that skill retention could be localized to a few layers in LLMs. We also find that zero-order gradients (Hinton, 2022; Sung et al., 2024) are promising and computationally cheaper alternatives to backpropagated gradients for pseudo-task clustering. Analysis of the skill-wise breakdown of performance reveals that language-only skills are the easiest to retain and improve, whereas multilingual multimodal skills exhibit significant forgetting in LiIT.

In summary, to the best of our knowledge, we are the first to explore the realistic setting of lifelong multimodal instruction tuning where the temporal stream of datasets may contain new skills, overlapping or rare tasks, and redundant samples. This scenario necessitates dynamic data selection strategies for efficient and effective learning. Our proposed method, LAMP, demonstrates superior retention as well as forward transfer of skills over time.

2 Related Work

Multimodal Instruction Tuning Datasets. While multimodal data, such as image-text pairs, has increased significantly, multimodal instruction-following data remains relatively scarce due to the time-consuming nature of human data collection. To address this, recent works (Liu et al., 2023b; Zhang et al., 2023a; Chen et al., 2023; Zhu et al., 2023; He et al., 2024a) have leveraged generative models to collect such data from existing image datasets. Li et al. (2023c) introduce the M3IT dataset with 2.4 million instances and 400 task instructions, translated into 80 languages. Xu et al. (2024) develop VISION-FLAN, a large-scale dataset of 187 tasks with expert-written instructions, ensuring diversity through iterative refinement. MultiInstruct (Xu et al., 2023) features 62 tasks across 10 categories, sourced from 21 open datasets.

Continual Instruction Tuning. In the era of multimodal LLMs, instructional datasets have raised several timely research problems. Therefore, it is crucial to develop sustainable models that can address emerging data and real-world challenges. Inspired by continual learning (Van de Ven & Tolias, 2019; Srinivasan et al., 2022; Lee et al., 2024b), a paradigm focused on enabling models to adapt to non-i.i.d., time-variant tasks, continual instruction tuning (CIT) (Chen et al., 2024; Zhang et al., 2023b) has recently been studied for (multimodal) LLMs that allows the model to adapt to multiple instruction tuning datasets sequentially without costly retraining. KPIG (He et al., 2024b) introduces a new CIT method that helps LLMs capture task-specific information and avoid overfitting general instructions by computing key-part information gain on masked parts to replay data and refine training dynamically. EProj (He et al., 2023) and Fwd-Prompt (Zheng et al., 2024) expand the CIT to the training of large multimodal models. EProj introduces new regularization and model expansion methods based on task correlations for continual instruction tuning of LMMs. Fwd-Prompt proposes a prompt-based approach that projects the prompt gradient into the residual space to minimize task interference while utilizing pre-trained knowledge, reducing negative forward transfer.

Data Selection. Data selection has been explored in the form of coreset selection in many works (Welling, 2009; Chen et al., 2010; Feldman et al., 2011). Uncertainty/loss/error-based methods estimate the difficulty of a sample from model confidence (Swayamdipta et al., 2020) or its training dynamics (Toneva et al.; Paul et al., 2021; Bachem et al., 2015). Zheng et al. (2023) address catastrophic accuracy drop at high pruning rates; Maharana et al. (2024) represent datasetd as undirected graphs and employ message passing to select the best subset. Gadre et al. (2024) investigate data selection for CLIP Radford et al. (2021) models. Evans et al. (2024) use learnability score (Mindermann et al., 2022) to accelerate training of CLIP models.

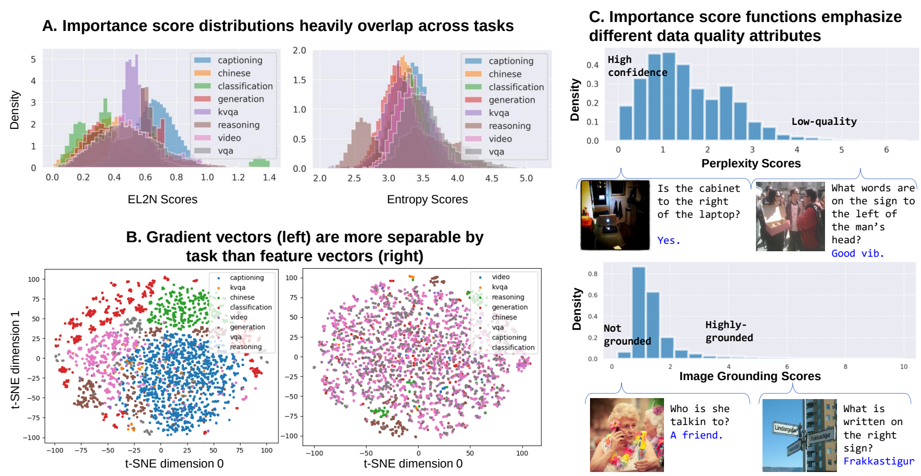

Figure 2: **A: Sample distributions for different visual language tasks** in M3IT (Li et al., 2023b) based on two importance scores, EL2N and entropy. **B: t-SNE visualization of sample vectors** based on their gradients and features. **C: Histogram** of Perplexity and Image Grounding scores. We visualize a few samples from M3IT with prompts (black) and ground-truth answers (blue).

3 Limitations of Score-based Data Selection in Lifelong Multimodal Instruction Tuning

3.1 Locality of the Sample Importance: Data Selection Depends on the Data

Score-based selection methods are widely used to assess the importance of training samples across modalities (Zheng et al., 2023; Liu et al., 2024; Marion et al., 2023; Evans et al., 2024; Gadre et al., 2024). We analyze two importance scores in multimodal instruction tuning: the EL2N score (Paul et al., 2021) and entropy score (Coleman et al.). EL2N measures the L2-norm of the output error vector, while entropy reflects the uncertainty in the output probabilities. Using the M3IT dataset (Li et al., 2023b), which includes eight tasks: captioning, multilingual (Chinese), classification, generation, knowledge VQA (kvqa), reasoning, videoQA, and VQA, we compute score distributions. As shown in Figure 2A, relying on a single importance score metric, such as EL2N or entropy, is insufficient to differentiate meaningful samples across a diverse range of tasks. For instance, selecting higher EL2N scores tends to favor for generalization (Paul et al., 2021), *captioning* samples over *kvqa*, leading to a skewed dataset.

In addition, the effectiveness of different importance scores varies based on the task and dataset at hand. The perplexity score is effective for filtering out low-quality samples in VQA that generally occur in the tail end of its distribution (see Figure 2C). Tasks such as *kvqa* (in purple) and *captioning* (in blue) are more separable via their entropy scores than EL2N scores. Moreover, Zheng et al. (2023); Swayamdipta et al. (2020) show that the most effective training subset contains a balanced mix of easy, difficult, and ambiguous samples. A biased score estimator may assign higher or lower scores to too many samples, making it hard to select the most effective subset. Thus, we need a different, generalizable strategy for assessing sample importance across multiple datasets during multimodal instruction tuning.

3.2 Importance of Vision-language Data with Image Grounding Score

The perplexity score measures the likelihood of a given sequence of tokens as predicted by an autoregressive generative model and has been used for selecting samples with higher instruction following difficulty in language datasets (Li et al., 2024). For MLLMs, the perplexity function additionally conditions on the input image tokens within the model. A lower perplexity score implies that the model assigns a higher probability to the output tokens. We compute the perplexity of a multimodal data instance $\boldsymbol{z}_i$ for an MLLM with weights $\boldsymbol{\theta}$ as:

$$\text{PPL}(\boldsymbol{z}_i) = \exp(\frac{1}{|\boldsymbol{z}_i|}\sum_{e_j \in \boldsymbol{z}_i} \text{NLL}(e_j)), \quad \text{where } \text{NLL}(e_j) = -\log(e_j | e_{<j}, \boldsymbol{I}; \boldsymbol{\theta}). \tag{1}$$

where $\text{NLL}(e_j)$ indicates the negative log-likelihood of token e_j in the sequence $\boldsymbol{z}_i$ comprising of image $\boldsymbol{I}$ and tokens e. As discussed in the previous section, perplexity is useful for detecting low-quality multimodal samples (see Figure 2C top). Motivated by the effectiveness and generalizability of the perplexity score (Marion et al., 2023; Li et al., 2024), we further modify this scoring function to distill the importance of the image in a multimodal data instance. We compute the image grounding score of a multimodal data instance $\boldsymbol{z}_i$ with image $\boldsymbol{I}$ and tokens e as:

$$\text{IG}(\boldsymbol{z}_i) = \frac{\text{PPL}(e)}{\text{PPL}(e, \boldsymbol{I})} \tag{2}$$

A higher IG score is assigned when the model assigns a higher probability to the text when conditioned on the image, compared to when the image is absent. Conversely, a lower IG score indicates that the image has little to no effect on the text's probability. As shown in Figure 2C bottom, an image-query pair with a higher IG score requires the model to carefully understand the visual scene (e.g., reading the text on a sign in the image). In contrast, examples where the model can predict the answer without seeing the images represent lower IG scores. Thus, the IG scoring function allows us to discard multimodal samples that do not leverage the multimodal functionality of MLLMs.

4 Lifelong Multimodal Instruction Tuning via Multi-way Data Selection

4.1 Problem Statement

This paper tackles the problem of LiIT over a sequence of multiple large datasets. Let $\mathcal{D}_0, ..., \mathcal{D}_{T-1}$ be a set of accessible datasets where $\mathcal{D}_t = \{\boldsymbol{x}_i^t, p_i^t\}_{i=1}^{N_t}$ denotes the dataset released in the timestep t, composed of N_i image-text pairs. Formally, we aim to train a multimodal model over multiple observed datasets for a given computational budget, such as FLOPs or training iterations. Given the model f parameterized by $\boldsymbol{\theta}$, the training objective at time step T is formulated as follows:

$$\arg\min_{\boldsymbol{\theta}} \frac{1}{T+1}\sum_{t=0}^{T}\sum_{i=0}^{\widehat{N}_t - 1} \mathcal{L}\left(f\left(\hat{\boldsymbol{x}}_i^t, \hat{p}_i^t; \boldsymbol{\theta}\right), \hat{y}_i^t\right) \quad \text{s.t.} \quad T \cdot (\widehat{N}_t - 1) \leq \tau, \tag{3}$$

where $\widehat{N}_t = r(N_t, T, \tau) \in \mathcal{N}$, $\widehat{N}_t \leq N_t$, and τ is the computational budget. $r(\cdot)$ denotes a decay function conditioned on T and τ, and $\hat{y}$ indicates the ground truth answer corresponding to the input data sample. Here, we constrain the minibatch iterations for multimodal instruction tuning per training timestep, by subsampling $\widehat{\mathcal{D}} = \{\hat{\boldsymbol{x}}_i, \hat{p}_i\}_{i=1}^{\widehat{N}}$, where $\widehat{\mathcal{D}} \subseteq \mathcal{D}, \quad \widehat{\mathcal{D}} \sim P(\widehat{\mathcal{D}} \mid \mathcal{D}, T, \tau)$. When a new multimodal instruction tuning dataset D_T is released, $\{\widehat{\mathcal{D}}_t\}_{t=0}^{T}$ is (re-) drawn for finetuning θ.

We propose the LAMP data selection method for the problem of lifelong multimodal instruction tuning and describe each of its steps in detail in the following sections.

4.2 Pseudo-task Clustering via Gradients

In the LiIT scenario, the model continuously updates its weights to incorporate new knowledge and refine its capabilities in specific tasks by training on unseen, meaningful data. The relative importance of each data sample evolves with changes in the model's state and expansion of the data pool at each time step. Therefore, adjusting the relative importance of samples within the data pool over time is crucial to faster and better optimization under restricted conditions. We accomplish this by first using gradient vectors from the model's current state to estimate skill clusters within the training data pool. As shown in Figure 2B and Figure 4, we find that gradient vectors are significantly more separable by skills than hidden state outputs of the model.

For a model with weights $\boldsymbol{\theta}_l$ for layer l, we compute the gradients of $\boldsymbol{\theta}_l$ for a sample $(\boldsymbol{z}_i, y_i) = ((\boldsymbol{x}_i, p_i), \hat{y}_i)$ using backpropagation. We obtain the data representation for the sample by concatenating weight gradient vectors $\nabla_{\boldsymbol{\theta}}\boldsymbol{z}_i = [\nabla_{\boldsymbol{\theta}_0}\boldsymbol{z}_i; \nabla_{\boldsymbol{\theta}_1}\boldsymbol{z}_i; ...; \nabla_{\boldsymbol{\theta}_{L-1}}\boldsymbol{z}_i]$, where L denotes the number of layers, and construct pseudo-task clusters in the data pool by performing k-means clustering over data representations of the seen and unseen datasets (see Figure 1). However, in practice, we find that not all gradients are necessary for distinguishing samples into multiple meaningful skills. Following Yoon et al. (2022), we cluster samples based on the gradients of essential layers, which leads to better performance compared to using gradients from all layers and is more memory-efficient as well (see discussion in Section 6).

4.3 Entropy-based Multi-way Data Selection

As discussed in Section 3.1, different scoring functions provide distinct advantages when selecting meaningful samples from a wide range of vision-language instruction tuning tasks, such as in-domain VQA, open-ended VQA, multilingual QA, visual grounding, commonsense reasoning, scientific/physical understanding, and OCR. To address the limitations of relying on a single data selection (pruning) method and promote collaborative decision-making when evaluating sample importance, based on the model's current state and data diversity, we propose a novel and versatile, multi-way data pruning strategy. First, we construct a function pool $\mathcal{S} = \{s_0(\cdot), ..., s_{N-1}(\cdot)\}$ where $s_n(\cdot)$ denotes n-th sample selection operation based on the corresponding importance score function. Here, we aim to identify the function $\hat{s}$ that maximizes the entropy of the distribution of scores over the samples of each pseudo-task cluster. For the k-th cluster $\mathcal{C}_k$ with m samples, we first extract the corresponding scores for each s_n using model weights θ at time step t. We approximate the distribution of scores, P_n^θ, by binning the range of normalized scores and calculating densities $\hat{p}_n^b$ over the $\mathcal{B}_n^{(k)}$ bins. Omitting cluster index k for brevity, the selection of $\hat{s}$ is formulated as:

$$\hat{s}^{(k)} = arg\max_{s_n} H\left(\hat{\boldsymbol{P}}_n^\theta\right), \quad \text{where } \hat{\boldsymbol{P}}_n^\theta = \{\hat{p}_n^b(x)\}, \ \forall \boldsymbol{b} \in \mathcal{B}^{(k)} \text{ and } \boldsymbol{x} \in \mathcal{C}_k. \tag{4}$$

A score function that yields higher average entropy in its distribution compared to other functions indicates a better ability to assess the uncertainty of the system (i.e., the model θ at timestep t). Moreover, we employ the CCS (Zheng et al., 2023) sampling strategy that aims to select a balanced representation of seen and unseen samples as well as frequent and rare tasks by sampling uniformly over the range of a score function. Hence, it is even more imperative to use a score function that is discriminative over its entire range. Since the resulting pseudo-task clusters may vary in size, we define a training budget T and divide it uniformly over the k clusters. The leftover data budget from clusters with sizes $|c_i|$ smaller than T/k are equally distributed over other clusters. This results in the selection of a skill-balanced subset from the pool of multi-task datasets in lifelong learning. Importantly, our proposed multi-way approach is highly flexible since the scoring function pool can be seamlessly extended with new scoring functions based on users' needs.

4.4 Combined Permanent-Temporal Data Pruning by Removing Redundancy

The steps of LAMP outlined in the previous sections enable the model to effectively select the most beneficial samples for each pseudo-task (i.e., skill) and train on them adaptively, based on the current model and evolving dataset distributions. However, as the data pool grows with the release of new instruction-tuning datasets, the computational burden of updating gradient-based sample vectors and ranking their importance increases. To keep the computational requirements manageable over time, we further implement a permanent data pruning strategy to iteratively reduce the size of the entire dataset pool. At the end of training at each timestep, we measure pairwise cosine similarities between samples within each cluster and prune those with maximum similarities, as they are highly redundant and contain overlapping knowledge. The similarity is computed in the semantic space which is well represented using hidden layer outputs of the model as shown in Figure 4 (Abbas et al., 2023; Sorscher et al., 2022). Larger clusters are prioritized for pruning to fit a predefined data pool budget D until all clusters retain a uniform number of samples. We refer to the version of LAMP with this combined permanent-temporal pruning as Lite-LAMP.

5 Experiments

5.1 Experimental Setup

Model. We conduct our experiments using the LLaVA 1.5 multimodal large language model (Liu et al., 2023a). It is trained on top of the Vicuna LLM (Chiang et al., 2023) using a corpus of approx-

Table 1: **Comparison of multimodal instruction tuning datasets.** Distribution of various skills in the datasets. RE: Referring Expression, OD: Object Detection, KD: Keypoint Detection.

Training Datasets	Dataset Size	Skills							
		VQA	Knowledge VQA	Captioning	Multi-lingual	Non-text Output	Video QA	Complex Reasoning	Language Only
LLaVA-1.5 (Liu et al., 2023b)	665K	✓	✓	✓	✗	RE	✗	✓	✓
M3IT (Li et al., 2023c)	2.1M	✓	✓	✓	✓	✗	✗	✗	✗
MiniGPT4 (Zhu et al., 2023)	3K	✗	✗	✓	✗	✗	✗	✗	✗
MANTIS (Jiang et al., 2024)	666K	✓	✗	✗	✗	✗	✗	✓	✗
LaMM (Yin et al., 2023)	250K	✓	✗	✓	✗	OD & KD	✗	✗	✗
VisionFLAN (Xu et al., 2024)	191K	✓	✓	✓	✗	✗	✗	✓	✗

imately 600K image-text caption pairs for pretaining of the vision projectors from the pre-trained CLIP visual encoder ViT-L/14 (Radford et al., 2021). Further, it is trained on the LLaVA instruction tuning dataset consisting of 665K text-only and vision-language instances. We adopt LoRA finetuning (Hu et al., 2021) of the LLaVA-1.5-7B model with the recommended hyperparameters.[1]

Datasets. For training, in addition to the LLaVA-665K instruction tuning dataset at $t = 0$, we consider the following order of datasets: M3IT (Li et al., 2023c), MiniGPT4 (Zhu et al., 2023), MANTIS (Jiang et al., 2024), LAMM (Yin et al., 2023) and VisionFLAN (Xu et al., 2024). Each dataset's temporal order, size, and skill composition are summarized in Table 1. We select standard evaluation datasets to measure performance on the skills enumerated in Table 1. These datasets and their corresponding task-specific evaluation metrics are listed in Table 4.

Metrics. We report various evaluation metrics from existing literature designed for understanding the continual learning phenomena in machine learning. The **Average Accuracy** (acc) at final timestep $t = T$ is the average of the model's performance across all skills (and across datasets within each skill). The **Relative Gain** (r) metric (Scialom et al., 2022) is the average of skill performances as a % of the respective upper bounds i.e., $r^T = \frac{1}{S}\sum_{s=1}^{S} \frac{P_s^t}{upper\,bound^s} \times 100\%$. We consider the best performances in each skill group in the sequential learning setting to be the upper bound in performances and report r for the final time step T. We also report the **Forgetting Rate**(f) which is the % drop in performance averaged across all skills and timesteps i.e., $f = \frac{1}{S \times T}\sum_{s=1}^{S}\sum_{t=1}^{T} \frac{min(P_s^t - P_s^{t-1}, 0)}{P_s^{t-1}} \times 100\%$.

LAMP Setup. The optimal value of k in the pseudo-task clustering step is computed from a grid search over values of k between 5 and 50, and selected absed on the WSS value of clusters.[2] In the score-based sample selection step, we use a bin size of 50 and discard the top and bottom 5% of samples for computing entropy as well as for CCS sampling, to remove outliers, low-quality samples, and uninformative data (Zheng et al., 2023). We use perplexity, image grounding (Section 3.2), EL2N (Paul et al., 2021) and entropy (Coleman et al.) score functions for $\mathcal{S}$ throughout the paper.

Baselines. We present baseline numbers on (1) Sequential and Multi-task training, (2) Random selection for experience replay (10% of past datasets), (3) Score-based selection methods, including Random, EL2N (Paul et al., 2021), Entropy (Coleman et al.), Perplexity (Marion et al., 2023), and (3) recent competitive data pruning baselines: SemDeDup (Abbas et al., 2023), Density-based Pruning (Abbas et al., 2024), and COINCIDE (Lee et al., 2024a). See Appendix for details.

5.2 Main Results

We present the main experimental results in Table 2 and show a breakdown of skill-wise accuracies at each time step for various methods in Figure 3. Our main findings are as follows:

Sequential training leads to catastrophic forgetting of skills. Multiple skills learned at t=0 (LLaVA) are forgotten at t=1,2 upon training the model on datasets containing a different set of skills. The M3IT dataset (t=1) does not contain grounding tasks and results in large drops in performance for the same. Similarly, the MiniGPT4 dataset (t=2) predominantly contains high-quality captions and causes forgetting of all other skills. The MANTIS dataset (t=3) improves performance on the MMMU dataset because it contains training instances with documents, charts, and visualization images. At t=4, the LAMM dataset contains object detection and keypoint detection tasks that

[1] https://github.com/haotian-liu/LLaVA/tree/main

[2] Within the sum of squares (WSS) is sum of squared distance between samples and their cluster centroids.

Table 2: **Overall results for lifelong multimodal instruction tuning.** Comparison of performance of LLaVA models trained on datasets selected using LAMP vs. other data selection methods.

Pruning Strategy	Data size at t	Relative Gain % (↑)	Forgetting Rate % (↓)	Avg. Acc. (↑)
Sequential	Full	68.0	26.0	32.0
Multi-task	Full	92.5	-	46.1
Random Experience Replay	Full	89.5	6.6	43.4
Random	25k	95.3	2.1	47.2
EL2N (Paul et al., 2021)	25k	82.4	12.2	43.9
Entropy	25k	79.6	15.1	41.6
Perplexity (Marion et al., 2023)	25k	91.4	9.6	45.2
Image-grounding (ours)	25k	92.3	5.6	45.6
SemDeDup (Abbas et al., 2023)	25k	76.4	6.4	38.5
Density-based (Abbas et al., 2024)	25k	78.1	5.1	39.6
COINCIDE (Lee et al., 2024a)	25k	89.5	3.9	44.7
LAMP	25k	102.3	0.9	50.5
LAMP	50k	107.2	0.2	51.7
LAMP	100k	**109.7**	**0.4**	**52.5**
LITE-LAMP	25k	99.7	1.3	49.6

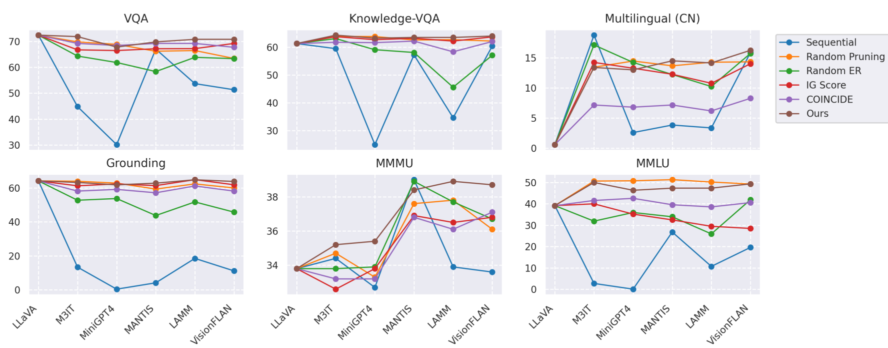

Figure 3: **Average accuracies per skill over time.** Comparison of average accuracies over time for each skill in our evaluation suite using various data selection methods. Higher is better.

promote recovery of the referring comprehension skill learned at $t = 0$, presumably due to a similarity in their non-textual output modalities. Finally, VisionFLAN (t=5) recovers performance on all skills except grounding and MMMU. Surprisingly, VisionFLAN also induces recovery of multilingual and unimodal skills (MMLU) despite not containing those tasks in its dataset composition. This method incurs nearly 26% forgetting across all timesteps and retains only 68% of its all-time high performances at the final timestep (see Table 2).

Random experience replay and pruning alleviate forgetting. Random experience replay using 10% data from past datasets significantly improves skills retention over sequential training, bringing down the forgetting rate from 26% to 6.6%. Random pruning of the combined set of past and incoming datasets results in better retention of skills with only a 2.1% forgetting rate using a fraction of the datasets for training as seen in Figure 3. This result demonstrates the need for applying data selection methods to the combined datasets rather than in isolation and serves as a strong baseline for lifelong multimodal instruction tuning. Notably, random pruning also improves the language-only skill of the underlying LLM in LLaVA. However, random pruning achieves a relative gain of 95%, falling short of reaching the best performance seen during sequential training.

LAMP minimizes forgetting and promotes forward transfer. LAMP outperforms all existing data selection methods for retaining and learning skills via instruction tuning over time. Score-based data selection methods generally fall short of random pruning because selecting samples from multi-task datasets based on scores leads to a skewed representation of tasks in the subset (see Figure 2). Our proposed scoring function, Image grounding, prioritizes the data samples where the outputs are strongly grounded in images (see Fig. 2C) and achieves relatively higher performance for multimodal tasks. SemDeDup (Abbas et al., 2023), DBP (Abbas et al., 2024), and COINCIDE (Lee et al., 2024a) rely on hidden layer outputs from the model to represent and prune data samples which can lead to imbalanced task distributions in the selected subset.

The use of gradient vectors in LAMP to pool samples into pseudo-task clusters before pruning ensures that all skills are well-represented at each time step resulting in the lowest forgetting rate i.e., 0.9%. Further, the multi-way score-based selection of *important* samples within those clusters

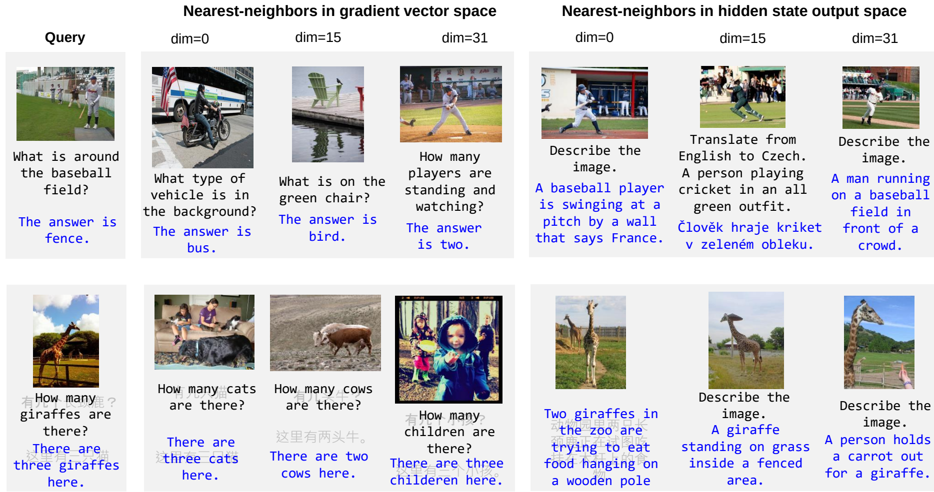

Figure 4: **Nearest neighbor samples** for query samples from VQA (top) and Chinese VQA (bottom) tasks in the gradient (left) and feature spaces (right).

Table 3: **Ablation results for LAMP**. Comparison of performance of the LAMP method to its ablated versions for lifelong multimodal instruction tuning.

Ablation Type	Values	Relative Gain % (↑)	Forgetting Rate % (↓)	Avg. Acc. (↑)
Within-cluster Pruning	*Multi-way*	**102.3**	**0.9**	**50.5**
	Image Grounding Score	96.3	2.7	48.8
	EL2N	97.4	1.5	49.1
Data Representations	*Gradients* (middle layer)	**102.3**	**0.9**	**50.5**
	Gradients (first layer)	97.2	1.8	47.9
	Gradients (last layer)	101.5	1.3	50.1
	Gradients (all layers)	98.9	2.5	49.1
	Hidden layer outputs (all layers)	96.5	4.3	47.4
Cluster Budget	*Uniform*	**102.3**	**0.9**	**50.5**
	Density-based	101.7	0.5	49.5

promotes forward transfer of skills and results in >100% relative gain unlike any other method in Table 2. The relative gain increases by nearly 5% on doubling the data budget from 25K to 50K samples and shows signs of plateauing with further increase in data.

Semantic deduplication for managing data complexity over time is effective. LITE-LAMP employs semantic deduplication of the training pool at the end of each timestep to reduce the computation complexity of extracting data representations at the next step. For the deduplication size of 100K samples (4x of training budget) across all timesteps, LITE-LAMP suffers minor drops in performance and forgetting as compared to LAMP.

6 ANALYSIS

Relative gains per skill. We present the skill-wise breakdown of relative gains of each method discussed in Table 2, in Figure 5A. The largest increases are observed for the language-only skill across all methods, except the image-grounding score which deprioritizes unimodal samples. This result is especially striking because none of the datasets in our experimental setting contain samples similar to those in the MMLU evaluation dataset. It suggests that the underlying LLM in LLaVA can recover this skill from similar multimodal samples. The next highest gains are seen for the knowledge VQA dataset, using our image-grounding score and the LAMP method. Multilingual skills appear to be the hardest skills to learn and retain over time, potentially because they utilize a different part of the model's vocabulary than other tasks and are included in the M3IT dataset only.

Multi-way pruning vs. single pruner. LAMP uses one among various importance score functions (or pruner experts) for each cluster to select a representative subset of importance samples from the cluster. This works better than using any one single metric across all clusters as shown in Table 3.

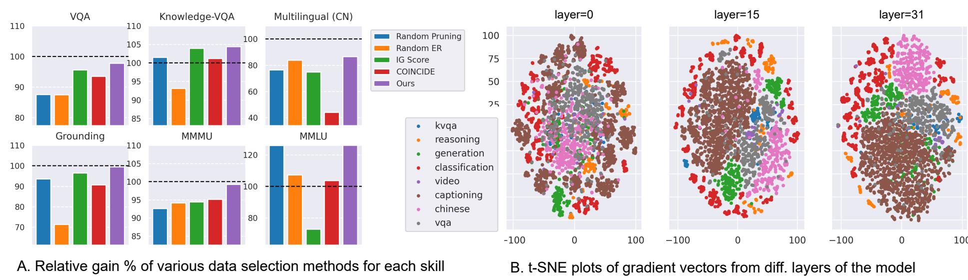

Figure 5: **A: Relative gain % for different skills** using various data selection methods in the lifelong multimodal instruction tuning setting. **B: t-SNE visualization of gradient vectors** of the M3IT dataset from layers of varying depth in the LLaVA model.

Importantly, this serves as a flexible framework for swapping or adding advanced 'expert' pruners, e.g., learnability score (Mindermann et al., 2022; Evans et al., 2024).

Source of data representations. Various data selection works propose different ways of representing data samples in a high-dimensional space. Methods designed for pruning pretraining datasets or single-task finetuning can effectively use semantic representations such as sentence embeddings (Abbas et al., 2023) and hidden state outputs (Sorscher et al., 2022; Maharana et al., 2024). However, we observe subpar performance with the use of semantic representations for pseudo-task clustering, as reported in Table 3. Gradient vectors are better representations of the skill component of a data sample as we show in Figure 4. Further, we experimented with gradients from all layers (similar to Xia et al. (2024); Liu et al. (2024)) as well as individual layers of the model. We observed the best performance with gradients from the middle layer only. Gradients from the last layer also work relatively well with LAMP but those from the first layer work poorly. This discrepancy correlates with the compactness of task clusters in the t-SNE plots of gradients from the corresponding layers, as demonstrated in Figure 5B.

Sampling budget across clusters. As outlined in Section 4, we sample a subset with an equal number of samples from each pseudo-task cluster in LAMP. The budgets leftover from smaller clusters are distributed equally across the remaining clusters. We experimented with more intuitive budgets for each cluster i.e., based on the density of the cluster members (Abbas et al., 2024). However, it did not result in significant changes in the performance.

Efficiency. Most data selection methods require extra steps to select an influential data subset and incur computational costs. LAMP expends additional memory and inference-time compute to effectively select data. We experiment with three methods to reduce this cost: (1) Zero-order gradients (Hinton, 2022), (2) Varying size of data pool in LITE-LAMP, and (3) gradients from a smaller model i.e., TinyLLaVA (Zhou et al., 2024). Using zero-order gradients instead of full gradients leads to a 2% drop in average accuracy and a 2.4% drop in relative gains. TinyLLaVA gradients demonstrate a similar drop, in addition to a higher forgetting rate. Conversely, the performance of LITE-LAMP improves with increasing size of the data pool post-deduplication (see Table 5).

7 CONCLUSION

Valuable visual instruction tuning datasets from various sources are released over time and often contain overlapping text-image pairs. To efficiently train lifelong adaptive MLLMs on these growing datasets, a scenario we call Lifelong Instruction Tuning (LiIT), we reformulate data selection so the model automatically selects meaningful samples from both old and new datasets, maintaining balance when incorporating new data. We observe that assessing sample informativeness with a static importance measure is challenging in LiIT, as it depends on the model's evolving capabilities and the shifting dataset distribution. To address this, we propose a scalable lifelong multimodal instruction tuning approach that dynamically balances sample efficiency and effectiveness through temporal multi-way data pruning. We show that training with samples selected by this method reduces catastrophic forgetting and enhances forward gain, using only a fraction of the original dataset, particularly for rare tasks with limited resources.

ETHICS STATEMENT

The intended use of LAMP is to enhance the vision-language instruction-following capabilities of MLLMs by training them on an integrated, ever-growing instructional dataset over time. This system does not pose any specific potential for misuse beyond the general risks associated with AI technology. However, instruction-tuned frameworks, including LAMP, are required to carefully consider the selection of training datasets and the intended purpose of usage to ensure the development of a reliable and trustworthy video-language inference system that supports stable, reliable, and safe AI.

REPRODUCIBILITY STATEMENT

This paper fully discloses all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions. To maximize reproducibility, we have included our code in the supplementary material. Also, we report all of our hyperparameter settings and model details in the Appendix.

REFERENCES

Amro Kamal Mohamed Abbas, Kushal Tirumala, Daniel Simig, Surya Ganguli, and Ari S Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.

Amro Kamal Mohamed Abbas, Evgenia Rusak, Kushal Tirumala, Wieland Brendel, Kamalika Chaudhuri, and Ari S Morcos. Effective pruning of web-scale datasets based on complexity of concept clusters. In *The Twelfth International Conference on Learning Representations*, 2024.

Olivier Bachem, Mario Lucic, and Andreas Krause. Coresets for nonparametric estimation-the case of dp-means. In *International Conference on Machine Learning*, pp. 209–217. PMLR, 2015.

Cheng Chen, Junchen Zhu, Xu Luo, Hengtao Shen, Lianli Gao, and Jingkuan Song. Coin: A benchmark of continual instruction tuning for multimodel large language model. *arXiv preprint arXiv:2403.08350*, 2024.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pp. 109–116, 2010.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations*.

Talfan Evans, Nikhil Parthasarathy, Hamza Merzic, and Olivier J Henaff. Data curation via joint example selection further accelerates multimodal learning. *arXiv preprint arXiv:2406.17711*, 2024.

Dan Feldman, Matthew Faulkner, and Andreas Krause. Scalable training of mixture models via coresets. *Advances in neural information processing systems*, 24, 2011.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

Yulu Gan, Sungwoo Park, Alexander Schubert, Anthony Philippakis, and Ahmed M Alaa. Instructcv: Instruction-tuned text-to-image diffusion models as vision generalists. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Jinghan He, Haiyun Guo, Ming Tang, and Jinqiao Wang. Continual instruction tuning for large multimodal models. *arXiv preprint arXiv:2311.16206*, 2023.

Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*, 2024a.

Yongquan He, Xuancheng Huang, Minghao Tang, Lingxun Meng, Xiang Li, Wei Lin, Wenyuan Zhang, and Yifu Gao. Don't half-listen: Capturing key-part information in continual instruction tuning. *arXiv preprint arXiv:2403.10056*, 2024b.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021a.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021b.

Geoffrey Hinton. The forward-forward algorithm: Some preliminary investigations. *arXiv preprint arXiv:2212.13345*, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 787–798, 2014.

Jaewoo Lee, Boyang Li, and Sung Ju Hwang. Concept-skill transferability-based data selection for large vision-language models. *arXiv preprint arXiv:2406.10995*, 2024a.

Jaewoo Lee, Jaehong Yoon, Wonjae Kim, Yunji Kim, and Sung Ju Hwang. Stella: Continual audio-video pre-training with spatiotemporal localized alignment. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023a.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. M3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023b.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. M3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023c.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7595–7628, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b.

Zikang Liu, Kun Zhou, Wayne Xin Zhao, Dawei Gao, Yaliang Li, and Ji-Rong Wen. Less is more: Data value estimation for visual instruction tuning. *arXiv preprint arXiv:2403.09559*, 2024.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Adyasha Maharana, Prateek Yadav, and Mohit Bansal. D2 pruning: Message passing for balancing diversity and difficulty in data pruning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. When less is more: Investigating data pruning for pretraining llms at scale. *arXiv preprint arXiv:2309.04564*, 2023.

Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, et al. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *International Conference on Machine Learning*, pp. 15630–15649. PMLR, 2022.

Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Mubarak Shah, and Fahad Khan. Pg-video-llava: Pixel grounding large video-language models. *arXiv preprint arXiv:2311.13435*, 2023.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale. In *International Conference on Machine Learning*, pp. 27074–27113. PMLR, 2023.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34:20596–20607, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pp. 146–162. Springer, 2022.

Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. Fine-tuned language models are continual learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6107–6122, 2022.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.

Tejas Srinivasan, Ting-Yun Chang, Leticia Pinto Alva, Georgios Chochlakis, Mohammad Rostami, and Jesse Thomason. Climb: A continual learning benchmark for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yi-lin Sung, Jaehong Yoon, and Mohit Bansal. Ecoflap: Efficient coarse-to-fine layer-wise pruning for vision-language models. In *International Conference on Learning Representations*, 2024.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9275–9293, 2020.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*.

Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

Max Welling. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1121–1128, 2009.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. In *Forty-first International Conference on Machine Learning*, 2024.

Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2023.

Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. Vision-flan: Scaling human-labeled tasks in visual instruction tuning. *arXiv preprint arXiv:2402.11690*, 2024.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Sk7KsfW0-.

Jaehong Yoon, Divyam Madaan, Eunho Yang, and Sung Ju Hwang. Online coreset selection for rehearsal-based continual learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Jaehong Yoon, Shoubin Yu, and Mohit Bansal. Raccoon: Remove, add, and change video content with auto-generated narratives. *arXiv preprint arXiv:2405.18406*, 2024.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 69–85. Springer, 2016.

Shoubin Yu, Jaehong Yoon, and Mohit Bansal. Crema: Multimodal compositional video reasoning via efficient modular adaptation and fusion. *arXiv preprint arXiv:2402.05889*, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023a.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model. https://llava-vl.github.io/blog/2024-04-30-llava-next-video/. 2024.

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad-Reza Namazi-Rad. Citb: A benchmark for continual instruction tuning. *arXiv preprint arXiv:2310.14510*, 2023b.

Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. Coverage-centric coreset selection for high pruning rates. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Junhao Zheng, Qianli Ma, Zhen Liu, Binquan Wu, and Huawen Feng. Beyond anti-forgetting: Multimodal continual instruction tuning with positive forward transfer. *arXiv preprint arXiv:2401.09181*, 2024.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*, 2024.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Table 4: **Evaluation datasets** used for measuring the performance of the trained model at each time step in our continual learning experiments.

Skill	Evaluation Datasets	Evaluation Metric
VQA	GQA (Hudson & Manning, 2019), LLaVA-Bench (Liu et al., 2023a)	Accuracy
Knowledge VQA	ScienceQA (Lu et al., 2022), OK-VQA (Marino et al., 2019), A-OK-VQA (Schwenk et al., 2022)	Accuracy
Multilingual (Chinese)	COCO-CN, Flickr-8K-CN (Li et al., 2023b)	BLEU Score
Grounding	RefCOCO, RefCOCO+, RefCOCOg (Kazemzadeh et al., 2014; Yu et al., 2016)	IOU
Reasoning	MMMU (Yue et al., 2024)	Accuracy
Language	MMLU (Hendrycks et al., 2021a;b)	Accuracy

Appendix

A Baselines

Sequential Training. In this method, the model is naively trained on the stream of instruction tuning datasets without any experience replay. Generally, this method sets a lower bound on the performance of the model on each evaluation task.

Multi-task Training. This method comprises pooling all of the datasets across time steps and training the model on this pooled dataset in one go. Generally, this method sets an upper bound on the performance of the model on various skill sets. However, if there are low-resource tasks in the dataset, it can lead to low performance on those tasks.

Coverage-based Coreset Selection (CCS). Zheng et al. (2023) introduces the method CCS where they divide a range of difficulty scores into equal-sized bins and randomly data samples from each bin with a uniform budget. This approach is motivated by maximizing coverage over the semantic space while ensuring an equal distribution of *easy* and *difficult* samples in the selected subset. Easy samples promote learning whereas difficult samples are information-dense and promote generalization. We use this method in LAMP for score-based selection.

Score-based Selection. Multiple importance score functions have been proposed over the years for various data types. We select the following for baseline experiments: (1) *Perplexity*: This metric is widely used for filtering language corpora (Marion et al., 2023), (2) *EL2N* (Paul et al., 2021): This metric is the L2-norm of the output error vector and is effective at low pruning ratios (or high retention rates), (3) Entropy (Coleman et al.): This score function is the entropy value of the output probability vector.

Embedding-based Selection. SemDeDup (Abbas et al., 2023) extracts semantic embeddings for pertaining datasets using a universal embedding transformer such as CLIP (Radford et al., 2021) for image-text pairs or Sentence Transformer[3] for natural language corpora and performs deduplication within k-means clusters. DBP (Abbas et al., 2024) assigns pruning budget to the clusters in SemDeDup using a cluster complexity score. COINCIDE (Lee et al., 2024a) clusters feature vectors into a large number of cluster e.g., k=10,000 to identify skill-concept clusters and samples non-uniformly from the clusters using a difficulty score metric.

[3] https://www.tensorflow.org/hub/tutorials/semantic_similarity_with_tf_hub_universal_encoder

Table 5: **Efficiency Results**. Comparison of performance of efficient versions of the LAMP method for lifelong multimodal instruction tuning. T = Training data size; $|D|$ = Size of data pool after permanent pruning in LITE-LAMP.

Method	**Relative Gain %** (↑)	**Forgetting Rate %** (↓)	**Avg. Acc.** (↑)
LAMP (T=25k)	102.3	0.9	50.5
+ TinyLLaVA gradients (Zhou et al., 2024)	99.6	2.9	48.1
+ Zero-order gradients (Hinton, 2022)	99.9	1.5	48.5
LITE-LAMP ($\|D\|$=100k)	99.7	1.3	49.6
LITE-LAMP ($\|D\|$=200k)	101.8	1.1	50.1
LITE-LAMP ($\|D\|$=500k)	102.5	0.7	50.8

B EXPERIMENTAL SETUP

Additional Details on LAMP Setup. We use random projections (Park et al., 2023; Xia et al., 2024) to reduce the dimensionality of gradient vectors extracted for the pseudo-task clustering step. We use a constant projection dimension of 8192 throughout our experiments.