DATA SELECTION FOR FINE-TUNING VISION LANGUAGE MODELS VIA CROSS MODAL ALIGNMENT TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Data-efficient learning aims to eliminate redundancy in large training datasets by training models on smaller subsets of the most informative examples. While data selection has been extensively explored for vision models and large language models (LLMs), it remains underexplored for Large Vision-Language Models (LVLMs). Notably, none of existing methods can outperform random selection at different subset sizes. In this work, we propose the first principled method for data-efficient instruction tuning of LVLMs. We prove that examples with similar cross-modal attention matrices during instruction tuning have similar gradients. Thus, they influence model parameters in a similar manner and convey the same information to the model during training. Building on this insight, we propose XMAS, which clusters examples based on the trajectories of the top singular values of their attention matrices obtained from fine-tuning a small proxy LVLM. By sampling a balanced subset from these clusters, XMAS effectively removes redundancy in large-scale LVLM training data. Extensive experiments show that XMAS can discard 50% of the LLaVA-665k dataset and 85% of the Vision-Flan dataset while fully preserving performance of LLaVA-1.5-7B on 10 downstream benchmarks and speeding up its training by 1.2×. This is 30% more data reduction compared to the best baseline for LLaVA-665k.

1 INTRODUCTION

Large Vision-Language Models (LVLMs) have demonstrated impressive capabilities in understanding and reasoning over multimodal inputs (Liu et al., 2023; 2024a; OpenAI, 2023). LVLMs need to be trained on large data to obtain satisfactory performance. However, the amount of information in large datasets does not scale linearly with their size, due to redundancy (Sorscher et al., 2022). This raises the following key question: *can we eliminate redundancy in large training datasets of LVLMs without harming their performance?* Answering this question enables efficient training and guides data collection.

There has been a lot of recent efforts in developing data-efficient methods for training foundation models. For Large Language Models (LLMs), heuristic metrics such as middle perplexity (Marion et al., 2023), high learnability (Zhou et al., 2023b), large gradient norm (El2N) Paul et al. (2021), and highest uncertainty (Bhatt et al., 2024; Maharana et al., 2023) are commonly used. Other heuristics remove duplicates (Abbas et al., 2023) or select central examples in the embedding space (Bhatt et al., 2024). For LVLMs, heuristics based on CLIP-Score (Gadre et al., 2023; Chen et al., 2024), influence function (Liu et al., 2023), or sampling from activation clusters of carefully chosen layers (Lee et al., 2024) have been proposed. Notably, none of existing methods outperform random selection at various subset sizes, as we confirm in our experiments.

In this work, we address this problem from an optimization perspective, by studying the effect of every example on minimizing the training loss. As machine learning models are trained with gradient methods, examples that have similar gradients during the training affect the model parameters in a similar manner.

Hence, redundancy for training should be defined as gradient similarity (Mirzasoleiman et al., 2020). However, identifying examples with similar gradients during training becomes very challenging for LVLMs. First, LVLMs have billions of parameters and gradient similarity in such a high-dimensional space becomes vacuous and prohibitively expensive to calculate. Besides, as gradients change during the training, one should take into account the similarity between high-dimensional gradients during the entire training process. Finally, image and text embeddings lie in different spaces, creating a gap between the modalities: the distances among image embeddings, text embeddings, and image–text embeddings have different magnitudes. This phenomenon has also been observed in prior work (Yi et al., 2024; Role et al., 2025). As a result, the part of the gradient (captured by attention) corresponding to cross-modal alignment has a different magnitude than the part corresponding to individual modalities. This makes similarity calculation based on full gradients ineffective.

In this work, we address the above challenges and propose a theoretically-rigorous and efficient method to eliminate redundancy in large training datasets of LVLMs. First, we analyze a single-layer transformer and prove that the pairwise gradient distance between examples at a checkpoint can be upper-bounded by the distance between their cross-modal attention matrices. Then, we show that for instruction tuning where the Hessian is small, examples that have bounded distance between their cross-modal attention matrices at two checkpoints have bounded gradient distance between the checkpoints. Finally, we propose Cross Modal Alignment SVD (XMAS) that fine-tunes a small proxy VLM and tracks the trajectory of largest singular values of cross-modal attention matrices of examples at a few checkpoints during the training. XMAS finds examples with similar gradients by clustering attention trajectories. Top singular values of a matrix are closely related to its norm. So, examples that have similar top singular values for their cross-modal attention matrices have similar gradient norms (i.e. amount of cross-modal alignment). XMAS finds clusters of examples with similar gradient norms throughout the entire training. Within every cluster, examples are learned together (at a similar pace) and thus have similar learning dynamics. Every cluster contains subgroups of examples with similar gradient vectors (directions) with slightly different alignment trajectory patterns. By sampling examples with the most stable alignment trajectory, XMAS selects the central example from every gradient subgroup. Selecting a balanced subset of the most stable examples from alignment trajectory clusters eliminates redundancy and ensures superior performance on any unseen downstream task. We also theoretically analyze the convergence of training on XMAS subsets.

Our experiments demonstrate that XMAS outperforms existing baselines at various data budgets and can discard 50% of the LLaVA-665k dataset and 85% of the Vision-Flan dataset while fully preserving performance of LLaVA-1.5-7B on 10 downstream benchmarks, and speeding up its instruction tuning (including the time for data selection) by 1.2×. This is 30% more data reduction compared to the best baseline for LLaVA-665k. We also conduct an extensive ablation study on different components of our method.

2 Related Works

High-quality data is crucial for ensuring satisfactory performance of LLMs and LVLMs.

Data-efficient Training of LLMs. For instruction tuning, manually crafted high-quality instruction/response pairs was shown highly effective (Zhou et al., 2023a). Motivated by this, several studies explored using LLMs such as ChatGPT, or training on textbooks (Eldan & Li, 2023; Li et al., 2023c; Chen et al., 2023). Metrics such as diversity (Bukharin & Zhao, 2023; Du et al., 2023; Tirumala et al., 2023), difficulty (Bhatt et al., 2024; Marion et al., 2023; Zhou et al., 2023a), middle perplexity rankings (Marion et al., 2023), high gradient norm (EL2N) (Paul et al., 2021), memorization ranking (Biderman et al., 2023), and high learnability (difference between initial and final loss values) (Zhou et al., 2023b) have been explored. However, these methods assign similar scores to similar examples and thus cannot eliminate redundancy. To address this, SemDeDup (Abbas et al., 2023) removes redundancy by clustering embeddings. D2 Pruning (Maharana et al., 2023) prunes data using graph-based message passing to balance diversity and difficulty.

Despite being effective for LLMs, these methods perform poorly for LVLMs and are often outperformed by random selection, as we will confirm in our experiments.

Data-efficient Training of LVLMs. There has been recent efforts for selecting high-quality multimodal data. CLIP-Score (Gadre et al., 2023) selects examples with highest image–text similarity based on a pre-trained CLIP model. However, CLIP-Score overlooks question–answer relevance and thus performs poorly for LVLMs. Self-Sup (Sorscher et al., 2022) clusters embeddings and selects examples closest to cluster centroids. SELF-FILTER trains a scoring model along with the VLM to learn difficulty of training instructions based on feature extracted from CLIP and GPT4V. Then, it uses the scoring model to select challenging instructions and filters them for diversity (Chen et al., 2024). TIVE (Liu et al., 2024c) selects examples that have the largest gradient similarity (influence) to other examples in the same task and selects more from tasks with smallest average influence. SELF-FILTER and TIVE are very expensive and yield suboptimal performance. Most recently, COINCIDE (Lee et al., 2024) proposed to train a proxy LVLM and cluster examples based on activations of carefully selected layers, and sampling more from clusters that are closer to each other and less from denser clusters (Lee et al., 2024). However, none of existing methods outperform random selection at various subset sizes, as we will confirm in our experiments.

Targeted Data Selection. Targeted data selection methods such as LESS (Xia et al., 2024) and ICONS (Wu et al., 2024) select influential training samples with largest gradient similarity to a validation set. Targeted data selection approaches have two major limitations: (i) they require computing gradients for every training example, which is even more computationally expensive than directly fine-tuning on the full dataset; and (ii) they rely on a validation set, which is often unavailable or impractical when training models intended for a broad range of downstream tasks. In our work, we do not assume access to a validation data.

3 Problem Formulation

Large Vision Language Model (LVLM). An LVLM consists of a vision encoder, an LLM, and a projector. We denote all the model parameters by ϕ_{all}. An input (v^i, t^i, y^i) to LVLM consists of an image v^i, an instruction t^i and the corresponding answer y^i. The encoded image is projected to the language space using the projector, before being concatenated with the instruction and fed into the language model. The response y^i is then sampled from the following conditional probability distribution:

$$p_\phi(y^i|v^i, t^i) = \prod_{j \in V} p_{\phi_{all}}(y^i_j|v^i, t^i, y^i_{<j}). \tag{1}$$

where y^i_j denotes the token at index j and $y^i_{<j}$ denotes all tokens before index j.

Visual Instruction Tuning (VIT). To adapt a pre-trained LVLM for following specialized task instructions, visual instruction tuning (VIT), which is a type of supervised fine-tuning (SFT), is employed on a dataset $\mathcal{D}_{\text{VIT}} = \{(v,t,y)^i\}_{i \in V}$. During instruction-tuning, the vision encoder is kept frozen and only the projector and the LLM are trained. We use ϕ to denote the trainable parameters. Therefore, the visual instruction tuning objective is to minimize the following negative log likelihood loss:

$$\min_{\phi} \mathcal{L}(\phi, \mathcal{D}_{\text{VIT}}) = -\frac{1}{|V|} \sum_{(v,t,y)^i \in \mathcal{D}_{\text{VIT}}} \left[\log p_\phi(y^i|v^i, t^i)\right] \tag{2}$$

In practice, gradient methods are applied to train the model by minimizing the above loss function.

Finding Redundant Clusters in the VIT Data. Consider an LVLM with parameters ϕ. Our goal is to find a clustering of the data $V = \{C_1 \cup \dots \cup C_K\}$ such that in every cluster C_k examples have similar gradients during the entire instruction tuning process. Formally, let Φ be the set of trainable parameters of the model during fine-tuning. Then we wish to find a solution to the following problem:

$$V = \{C_1 \cup \dots \cup C_K\} \quad \text{s.t.} \quad \max_{\phi \in \Phi} \|\nabla \mathcal{L}_i(\phi) - \nabla \mathcal{L}_j(\phi)\| \leq R \quad \forall i,j \in C_k, \forall k \in [K], \tag{3}$$

where $\nabla\mathcal{L}_i(\phi)$ is the gradient of example i at parameter ϕ. Examples in every cluster have similar gradients during instruction tuning and hence are redundant w.r.t each other for training.

Sampling a Balanced Subset From Clusters. Having the above clustering, for a given data budget B, we sample a balanced (i.e., same number of examples from each cluster) subset of examples $S \subseteq V$ from all the clusters $\{C_1 \cup \cdots \cup C_K\}$ to train the target model. In doing so, we effectively eliminate redundancy in the multimodal data, and ensure satisfactory performance on various (unseen) downstream tasks.

4 Method: Extracting Non-redundant Subsets of the VIT Data

Solving Eq. 3 is very challenging as it requires training the model, saving the gradients of all the training examples after every parameter update, and clustering the concatenated gradient vectors. However, for LVLMs with billions of parameters, this becomes infeasible and does not improve the training efficiency.

To address this, we wish to train a small proxy VLM and use its training dynamics to approximate the pairwise gradient distances during instruction tuning. If this can be done, one can cluster the training examples based on the estimated gradient distances obtained from the proxy, and randomly sample a balanced subset from the clusters to train the larger target LVLM on the non-redundant subset. If the proxy model is small-enough, training the proxy to find the non-redundant subset and instruction tuning the larger target LVLM will be faster than instruction tuning on the full data.

4.1 Finding Clusters of Examples with Similar Gradient

In this section, we wish to answer the following question: when fine-tuning a proxy VLM on the instruction-tuning data, which statistics can be used to upper-bound pairwise gradient distances during the training?

Answering the above question requires understanding the mechanism by which VLMs learn from the instruction-tuning data. Intuitively, instruction tuning aligns the vision and language modalities and enables the LLM to understand the content of the images. This is primarily done via the trainable attention matrices in the LLM structure. Formally, consider the l-th layer of the language decoder of a VLM with parameter ϕ with hidden dimension size D. The per-layer attention matrix for example $i \in V$ is defined as:

$$A_i^l(\phi) = \text{softmax}\left(\frac{Q_l \otimes K_l^T}{\sqrt{D}}\right) \in \mathbb{R}^{N \times N}, \tag{4}$$

where $N = n_I + n_T$ which is the total number of image and text tokens, and $Q_l, K_l \in \mathbb{R}^{N \times D}$ are the concatenated query and key matrices along the hidden dimension across all the attention heads. $A_i^l(\phi)$ consists of both the cross-modal attention and intra-modal attention terms. The cross-modal attention part $\chi_i^l(\phi) \in \mathbb{R}^{n_T \times n_I}$ corresponds to the bottom left block of of the per-layer attention matrix $A_i^l(\phi)$.

Definition 4.1 (Cross-modal alignment score σ)**.** *For data point i, we define $\sigma_i(\phi)$ as the sum of the top five singular values of its cross-modal attention matrix $\chi_i(\phi) = \sum_{l=1}^{L} \chi_i^l(\phi)$. Intuitively, σ captures the amount of cross-modal alignment for individual examples.*

Remark. Cross-modal attention matrices are transferrable between proxy and target LVLMs (Zhao et al., 2024). This implies that cross-modal alignment scores obtained based on a proxy VLM closely estimates alignment for the larger LVLM. We will empirically confirm this observation in our ablation studies.

4.1.1 Bounding Gradient Distance via Attention Distance

Next, we analyze a single-layer transformer with one attention head and RMS layer normalization, trained using the Frobenius norm squared loss. This setting has been studied in several recent theoretical analysis of transformers (Ormaniec et al., 2024; Song et al., 2024). RMS layer normalization is standard practice in many recent open-source models, such as LLaMA and Mistral. In practice, the gain parameter g of RMS normalization is often initialized to a small value to help stabilize early training, thereby preventing activation blowups and facilitating stable learning in residual architectures Zhang & Sennrich (2019).

The following theorem shows that at every step t during the training, pairwise attention distances can be used to upper-bound pairwise gradient distances.

Theorem 4.1. *Consider a single-layer transformer with a single attention head with RMS layer normalization, that is trained using the Frobenius norm squared loss. Let D be the hidden dimensionality of the model, N be the number of input tokens, and $c \geq \|\phi^t\|$ be the upper-bound on the norm of model parameters. Then, if the gain g of RMS normalization layer is small enough $g < N^{-5/8}D^{-1/8}c^{-3/4}$ and for all examples $p \in V$ the distance between cross-modal attention matrices of the proxy and target models are bounded, i.e., $\|\chi_p(\phi^t_{proxy}) - \chi_p(\phi^t_{target})\|_F \leq T$, then pairwise attention distances of the proxy model $K^t_{ij} = \|\chi_i(\phi^t_{proxy}) - \chi_j(\phi^t_{proxy})\|_F$ for examples $i, j \in V$ dominate the bound on their pairwise gradient distance of the target model at every step t in training:*

$$\|\nabla\mathcal{L}_i(\phi^t_{target}) - \nabla\mathcal{L}_j(\phi^t_{target})\|_F \leq \frac{4}{\sqrt{3}} \cdot (K^t_{ij} + 2T) + \frac{8\sqrt{N}}{3\sqrt{3}c} = \Delta^t_{ij} \tag{5}$$

All the proofs can be found in Appendix A.

Remark. Under the assumptions of Theorem 4.1, we have $K^t_{ij} \leq 2\sqrt{N}$. Thus, for a good proxy model with small T, i.e. $T \ll \sqrt{N}$, K^t_{ij} dominates the upper-bound on pairwise gradient distances of the target model. Hence, clustering based on K^t_{ij} is similar (up to some error) to clustering based on gradients of the target model at step t.

4.1.2 Bounding Gradient Distance Throughout Finetuning

Theorem 4.1 shows that examples with similar cross-modal alignment score at a particular step t during training have similar gradients at step t. However, our goal is to find groups of examples with similar gradient *throughout* the training. Next, we show that for fine-tuning where loss has a small bounded curvature (Gekhman et al., 2024; Yang et al., 2024), examples that have similar cross-modal attention matrices two checkpoints also have similar gradients *between* those checkpoints.

Theorem 4.2. *Under the assumptions of Theorem 4.1, Suppose the per-example loss during fine-tuning admits a second-order Taylor approximation with bounded curvature, i.e., $\|\nabla^2\mathcal{L}_i(\phi^t_{target})\| \leq \beta \ \forall t$, Then, for any two checkpoints ϕ^{t_1}, ϕ^{t_2} where $\|\phi^{t_1} - \phi^{t_2}\| \leq \delta$, the largest pairwise attention distance between them provides an upper bound on the gradient distance at any intermediate checkpoint. Specifically, for all $t_z \in [t_1, t_2]$, we have:*

$$\|\nabla\mathcal{L}_i(\phi^{t_z}_{target}) - \nabla\mathcal{L}_j(\phi^{t_z}_{target})\|_F \leq \max\{\Delta^{t_1}_{ij}, \Delta^{t_2}_{ij}\} + 2\delta\beta = \Delta_2 \tag{6}$$

Remark. The above theorem implies that if two examples have similar attention matrices at a few checkpoints during instruction tuning of the proxy model, then examples have similar target gradients *throughout* instruction tuning. Since curvature β is small during fine-tuning, $\max\{\Delta^{t_1}_{ij}, \Delta^{t_2}_{ij}\} \leq \frac{8\sqrt{N}}{\sqrt{3}}$ will dominates the upper-bound of gradient distance in Eq 6. Based on this, we define alignment trajectory:

Definition 4.2 (Alignment trajectory.). *Consider r checkpoints $\{\phi^{t_1}, \phi^{t_2}, \ldots, \phi^{t_r}\}$ during fine-tuning a proxy model on $\mathcal{D}_{VIT}$. The cross-modal alignment trajectory of example $i \in V$ is defined as:*

$$T_i = \{\sigma_i(\phi^{t_1}), \sigma_i(\phi^{t_2}), \ldots, \sigma_i(\phi^{t_r})\} \tag{7}$$

where $\sigma_i(\phi^{t_j})$ represents the cross-modal alignment score of example i at checkpoint ϕ^{t_j}.

Algorithm 1 Data Selection Based on Cross-Modal Alignment Trajectories (XMAS)

```
Require: Training dataset $\mathcal{D}_{\text{VIT}}$, a fixed data budget $B$, number of clusters $K$
Ensure: Subset $S \subseteq \mathcal{D}_{\text{VIT}}$, $|S| \leq B$
$S \leftarrow \emptyset$
Train a small proxy VLM and cluster examples in $\mathcal{D}_{\text{VIT}}$ based on their alignment trajectories
Compute Instability score $S_i$ for each example in $\mathcal{D}_{\text{VIT}}$ according to Eq 8
Sort clusters by size to get $C = \{C_1, C_2, \dots, C_K\}$
for $k = 1$ to $K$ do
    if $|C_k| \leq R_k = \frac{B-|S|}{K-k+1}$ then
        $S \leftarrow S \cup C_k$
    else
        $S \leftarrow S \cup C'_k$ where $C'_k \subset C_k$ is the subset of $R_k$ most stable examples
return $S$
```

Clustering alignment trajectories. Having cross-modal alignment trajectories for all examples in the dataset using the proxy model, we cluster these trajectories using the K-means clustering. In doing so, we get clusters of examples with similar gradients during the training $\{C_1, C_2, \dots, C_K\}$.

4.2 Sampling a Balanced Subset from Gradient Clusters

Next, we sample a balanced subset from the alignment trajectory clusters to eliminate redundancy. While random sampling from the clusters is already effective, sampling examples with more stable (less oscillating) trajectories yields a better performance in practice.

Definition 4.3 (Instability score). *The instability score of example $i \in V$ is the total oscillation in the cross-modal alignment score of i during fine-tuning:*

$$S_i = \sum_{j=1}^{T} |\sigma_i(\phi^{t_j}) - \sigma_i(\phi^{t_{j-1}})|. \tag{8}$$

Intuitively, examples with smallest instability score within every cluster are centers of subgroups in that cluster. Sampling examples with smallest instability score ensures selecting a diverse set of representative examples from the clusters. We will confirm the effectiveness of stability sampling in our ablation studies.

4.3 Data Selection Based on Cross Modal Alignment Trajectories (XMAS)

To summarize, our method, XMAS includes three main steps. First, we fine-tune a proxy model to get the cross-modal alignment trajectories (see Def. 4.2) for all the examples. Then, we cluster these trajectories to find examples with bounded gradient difference throughout instruction tuning. Finally, we sample a balanced subset of stable points from these clusters. The pseudocode of XMAS is given in Algorithm 1 and a visualization can be found in Figure 5 in the Appendix.

Convergence of XMAS. Finally, we analyze the convergence of finetuning target model on the subset. The following corollary shows that training on the subset selected by XMAS converges to a neighborhood of the optimal solution found by training the target LVLM on the full dataset. For brevity, we show the informal version of the corollary here and defer the formal statement of the corollary and its proof to Appendix A.3.

Corollary 4.3 (Informal: Convergence of XMAS). *Under the assumptions of Theorem 4.2, applying incremental gradient methods with stepsize η on subsets found by XMAS, converges to a neighborhood of the solution ϕ^* found by training on full data.*

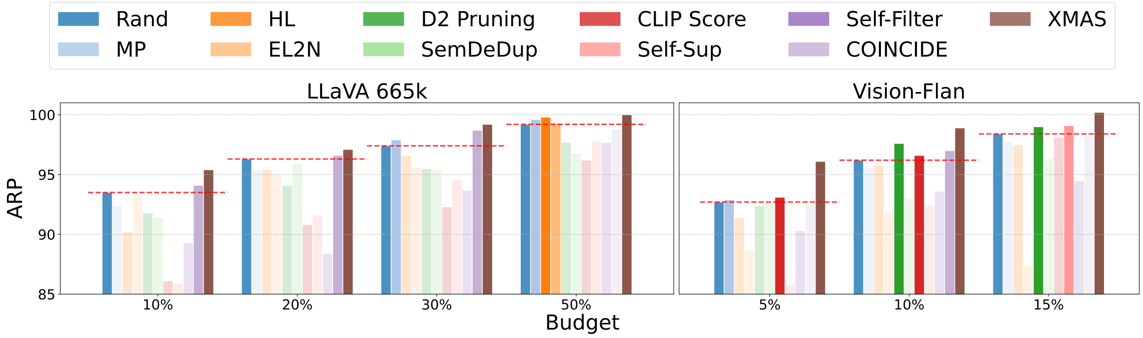


Figure 1: The average relative performance (ARP) of different subsets of (left) LLaVA 665k and (right) Vision-Flan when fine-tuning LLaVA-1.5-7B. Methods that outperform random selection are shown in opaque. XMAS is the only method that surpasses random selection across different budgets on both datasets. MP and HL finetune the target model on full data to find subsets and thus do not yield any speedup.

5 Experiments

In this section, we first describe our experimental settings, followed by an empirical evaluation of our method on two widely used VIT datasets in Section 5.1. We then report run time of XMAS and analyze the impact of different design choices (proxy models, number of checkpoints, number of clusters, sampling strategies, and attention matrices) in our approach. Additional results and qualitative analysis are deferred to Appendix C.

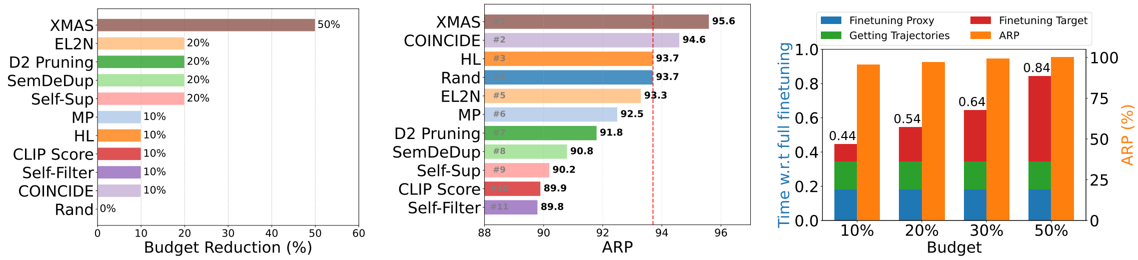
Models. For the target LVLMs, we use the pre-trained LLaVA-1.5-7B model (Liu et al., 2024a). For the proxy model, we use the TinyLLaVA-2B (Zhou et al., 2024).

VIT datasets. We apply coreset selection to two separate VIT datasets: LLaVA 665k (Liu et al., 2024a) and Vision-Flan (Xu et al., 2024). LLaVA 665k comprises 665k VIT examples collected from 12 vision-language datasets. On the other hand, Vision-Flan consists of 191 vision-language tasks, each featuring approximately 1k expert-labeled VIT samples, amounting to a total of 186k instances.

Training details. In all experiments, we fine-tune the target models using LoRA (Hu et al., 2022) for one epoch, following the official finetuning hyperparameters specified in LLaVA-1.5. For proxy models, we train without LoRA for one epoch, following the official hyperparameters specified in TinyLLaVA. This results in a total of $T = 7$ checkpoints. For K-means, we set the number of clusters K to 1000.

Baselines. We compare XMAS against 11 data selection baselines, including Random selection (**Rand**), Middle Perplexity (**MP**) (Marion et al., 2023), and Highest Learnability (**HL**) (Zhou et al., 2023b). In addition, we consider EL2N (Paul et al., 2021), Self-Sup (Sorscher et al., 2022), CLIP-Score (Hessel et al., 2021), SemDeDup (Abbas et al., 2023), D2 Pruning (Maharana et al., 2023). We also include two recent methods proposed for LVLMs: Self-Filter (Chen et al., 2024) and COINCIDE (Lee et al., 2024).

Evaluation. We evaluate the performance of the fine-tuned target models on reasoning, hallucinations, perception, and cognition capabilities. For all the experiments, we evaluate the aforementioned capabilities of the fine-tuned model on POPE (Li et al., 2023a), TextVQA (Singh et al., 2019), MME-Perception (Liang et al., 2024), ScienceQA (Lu et al., 2022), VizWiz (Gurari et al., 2018), MMBench (Liu et al., 2024b), LLaVABench (Liu et al., 2023), MMVet (Yu et al., 2023), VQAv2 (Goyal et al., 2017) and GQA (Hudson & Manning, 2019) datasets. We follow the same evaluation protocols outlined in LLaVA-1.5. We measure the relative performance of subsets as (subset performance / full data performance) × 100%. To compare between different methods, we **A**verage the **R**elative **P**erformance (ARP) across all evaluation datasets.



(a) Data reduction to 100%

(b) LLaVA-13B

(c) XMAS Time and ARP

Figure 2: (left) Data reduction to reach 100% Average relative performance (ARP) of LLaVA-1.5-7B on LLaVA 665k. XMAS obtains 30% more data reduction over the best baselines. (middle) ARP ranking of different 10% subsets of LLaVA 665k when fine-tuning LLaVA-1.5-13B. (right) ARP and ratio of total time (selection + training) w.r.t training target model on full dataset for XMAS at different budgets on LLaVA 665k. XMAS reduces training time by a factor of 0.84 (1.2× speedup) to reach 100% ARP.

5.1 Main results

LLaVA 665k dataset. As shown in Fig. 1 left, XMAS outperforms all baselines across different budgets. Relying on a single metric is worse than random selection for small budgets of 10-30% and worse than COINCIDE. This set of experiments confirms the observation made by (Lee et al., 2024) that single metrics lead to biased and redundant selection that, in turn, leads to sub-optimal results. Notably, XMAS reaches the performance of fine-tuning on the full datasets at only 50%, which is 2x data reduction compared to fine-tuning on the full dataset. This is 30% more data reduction over the best baselines as shown in Fig. 2a. When training LLaVA-13B on the subsets, Fig. 2b demonstrates that XMAS still achieves the highest ARP, demonstrating its effectiveness for training larger LVLMs.

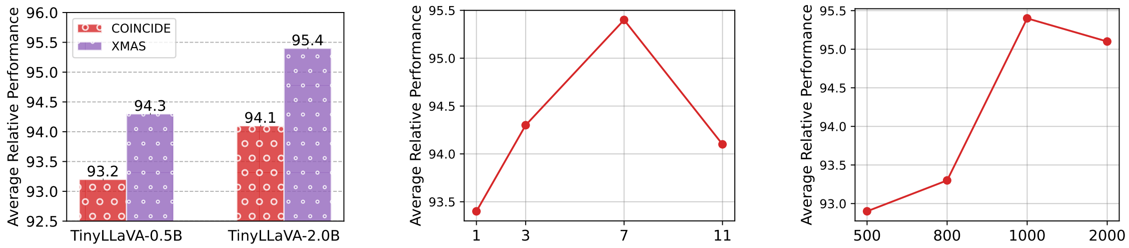
Vision-Flan dataset. Fig. 1 right demonstrates the superior performance of XMAS, consistently outperforming random sampling by at least 2% while other baselines fail to surpass random at all budgets. Our method matches the full performance with only 15% of the total samples, which is 6.7x data reduction compared to fine-tuning on the full dataset.

5.2 Ablation studies and Analysis

Computation time. Fig. 2c reports the total time for data selection (proxy fine-tuning + trajectory extraction) and target model fine-tuning on LLaVA-665k. The clustering and sampling steps in our method incur negligible cost. We observe that XMAS at 50% data retains full-dataset performance while delivering a 1.2× speedup. Although COINCIDE's feature extraction is faster, it can only discard 10% of the data, resulting in higher fine-tuning costs for the target model and making it 1.1× slower than full training.

Different proxy models. To investigate the effect of the proxy model, we try two different scales of TinyLLaVA which are 0.5B and 2B. Fig. 3a compares the performance of COINCIDE and XMAS when varying the proxy models. XMAS outperforms COINCIDE for different proxy models.

Number of checkpoints. Fig. 3b illustrates the performance of XMAS for varying number of proxy checkpoints. Using 7 checkpoints covering one pass of the full dataset yields the best performance of 95.4%. While using only 1(3) checkpoint reduces the computation cost, it harms the performance by roughly 2(1)%. Notably, increasing the number of checkpoints to 11 not only increases the computation cost but also decreases the relative performance to 94.1%.



(a) Proxy models

(b) Number of checkpoints

(c) Number of clusters

Figure 3: Average performance relative to full data for 10% subsets of LLaVA 665k found by COINCIDE and XMAS when varying (left) proxy model, (middle) number of checkpoints and (right) number of clusters.

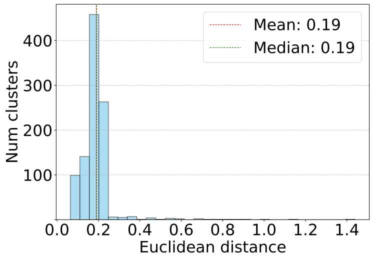


Figure 4: Per-cluster Euclidean distance between the cross-modal alignment trajectories of proxy model and target model on LLaVA 665k. Distance between trajectories of proxy and target models is very small.

Table 1: Average relative performances (ARP) over full data when training LLaVa-1.5-7B on 10% subsets of LLaVA 665K found by XMAS when using different (left) cluster sampling strategies and (b) attention matrices.

Strategy	ARP
Random	94.1
Instability	**95.4**

(a) Cluster sampling

Attn matrix	ARP
Full	94.0
Cross-modal	**95.4**

(b) Attention matrix

Number of clusters. Fig.3c illustrates the performance of XMAS for different numbers of clusters (K). Using 1,000 clusters yields the best performance of 95.4%. Increasing the number of clusters to 2,000 slightly lowers performance to 95.1% while fewer clusters (500 or 800) reduce performance to 92.9% and 93.3%.

Proxy vs target trajectories. Fig. 4 illustrates the per-cluster Euclidean distance between the alignment trajectories of proxy model (TinyLLaVA-2B) and target model (LLaVA-1.5-7B) on LLaVA 665k, i.e., $\sum_{i \in C_k} \|T_i^{\text{proxy}} - T_i^{\text{target}}\|_2 / |C_k|$. We see clearly that the trajectories of proxy and target models are similar as indicated by small Euclidean distance (< 0.25) for most of the clusters.

Cluster sampling strategy. Table 1a illustrates that cluster sampling based on the instability score performs better than random sampling by 1.3%. This validates our intuition in 4.2, which shows that sampling based on instability scores guides XMAS to select more representative examples.

Choice of attention matrices. Next, we study the effect of using the full vs cross-modal attention matrices. Table 1b shows that using only the cross-modal terms works better than using both intra-modal and cross-modal terms. This confirms that the cross-modal attention provides more informative signals than intra-modal for multi-modal data selection.

Additional ablation studies on the choice of alignment score, layer aggregation strategy, instability score, attention layers, and number of singular values are in Appendix C. Moreover, we show that COINCIDE is sensitive to the choice of layers used for feature extraction. Qualitative results of XMAS clusters are given in Fig. 7-10.

6 Conclusion

In this work, we present XMAS, an effective data selection method to improve data efficiency of visual instruction tuning for Large Vision Language Models (VLMs). By clustering examples based on their cross-modal alignment trajectories and sampling a balanced set of examples with the most stable trajectories from all the clusters, XMAS results in a significant reduction of the required training data without compromising the performance as compared to training on the complete dataset. Empirically, XMAS can drop 50% of LLaVA 665K and 85% of VisionFLAN datasets and train LLaVA-1.5-7B to the same accuracy of full data.

References

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. arXiv preprint arXiv:2303.09540, 2023.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. arXiv preprint arXiv:2412.08905, 2024.

Gantavya Bhatt, Yifang Chen, Arnav M Das, Jifan Zhang, Sang T Truong, Stephen Mussmann, Yinglun Zhu, Jeffrey Bilmes, Simon S Du, Kevin Jamieson, et al. An experimental design framework for label-efficient supervised finetuning of large language models. arXiv preprint arXiv:2401.06692, 2024.

Stella Biderman, Usvsn Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. Advances in Neural Information Processing Systems, 36:28072–28090, 2023.

Alexander Bukharin and Tuo Zhao. Data diversity matters for robust instruction tuning. arXiv preprint arXiv:2311.14736, 2023.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. Alpagasus: Training a better alpaca with fewer data. arXiv preprint arXiv:2307.08701, 2023.

Ruibo Chen, Yihan Wu, Lichang Chen, Guodong Liu, Qi He, Tianyi Xiong, Chenxi Liu, Junfeng Guo, and Heng Huang. Your vision-language model itself is a strong filter: Towards high-quality instruction tuning with data selection. arXiv preprint arXiv:2402.12501, 2024.

Qianlong Du, Chengqing Zong, and Jiajun Zhang. Mods: Model-oriented data selection for instruction tuning. arXiv preprint arXiv:2311.15653, 2023.

Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english? arXiv preprint arXiv:2305.07759, 2023.

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, 36:27092–27112, 2023.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning llms on new knowledge encourage hallucinations? arXiv preprint arXiv:2405.05904, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6904–6913, 2017.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3608–3617, 2018.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. ICLR, 1(2):3, 2022.

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.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3):535–547, 2019.

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

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355, 2023a.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In International conference on machine learning, pp. 19565–19594. PMLR, 2023b.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. arXiv preprint arXiv:2309.05463, 2023c.

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering, pp. 405–409, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26296–26306, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In European conference on computer vision, pp. 216–233. Springer, 2024b.

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

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. arXiv preprint arXiv:2310.07931, 2023.

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.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In International Conference on Machine Learning, pp. 6950–6960. PMLR, 2020.

OpenAI. GPT-4, 2023. Software available from `https://openai.com/`.

Weronika Ormaniec, Felix Dangel, and Sidak Pal Singh. What does it mean to be a transformer? insights from a theoretical hessian analysis. arXiv preprint arXiv:2410.10986, 2024.

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.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 3505–3506, 2020.

François Role, Sébastien Meyer, and Victor Amblard. Fill the gap: Quantifying and reducing the modality gap in image-text representation learning. arXiv preprint arXiv:2505.03703, 2025.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8317–8326, 2019.

Bingqing Song, Boran Han, Shuai Zhang, Jie Ding, and Mingyi Hong. Unraveling the gradient descent dynamics of transformers. Advances in Neural Information Processing Systems, 37:92317–92351, 2024.

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.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari Morcos. D4: Improving llm pretraining via document de-duplication and diversification. Advances in Neural Information Processing Systems, 36: 53983–53995, 2023.

Xindi Wu, Mengzhou Xia, Rulin Shao, Zhiwei Deng, Pang Wei Koh, and Olga Russakovsky. Icons: Influence consensus for vision-language data selection. arXiv preprint arXiv:2501.00654, 2024.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. arXiv preprint arXiv:2402.04333, 2024.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In International conference on machine learning, pp. 10524–10533. PMLR, 2020.

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.

Yu Yang, Siddhartha Mishra, Jeffrey Chiang, and Baharan Mirzasoleiman. Smalltolarge (s2l): Scalable data selection for fine-tuning large language models by summarizing training trajectories of small models. Advances in Neural Information Processing Systems, 37:83465–83496, 2024.

Chao Yi, Yuhang He, De-Chuan Zhan, and Han-Jia Ye. Bridge the modality and capability gaps in vision-language model selection. Advances in Neural Information Processing Systems, 37:34429–34452, 2024.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490, 2023.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. Advances in neural information processing systems, 32, 2019.

Xiangxiong Zhang. Notes for optimization algorithms spring 2023. 2023.

Wangbo Zhao, Yizeng Han, Jiasheng Tang, Zhikai Li, Yibing Song, Kai Wang, Zhangyang Wang, and Yang You. A stitch in time saves nine: Small vlm is a precise guidance for accelerating large vlms. arXiv preprint arXiv:2412.03324, 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.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. Advances in Neural Information Processing Systems, 36:55006–55021, 2023a.

Haotian Zhou, Tingkai Liu, Qianli Ma, Jianbo Yuan, Pengfei Liu, Yang You, and Hongxia Yang. Lobass: Gauging learnability in supervised fine-tuning data. arXiv preprint arXiv:2310.13008, 2023b.

A Proofs

Notations. Consider a data set with $|\mathcal{D}_{\text{VIT}}|$ samples where each sample consists of $N = n_I + n_T$ number of tokens with embedding dimension D. We denote the dataset as $(X_i, y_i)_{i=1}^{\mathcal{D}_{\text{VIT}}}$, where $X_i \in \mathbb{R}^{N \times d}$, and $y_i \in \mathbb{R}^N$ is the label of the dataset. Following the settings in Ormaniec et al. (2024), we consider a single-layer attention Transformer model with a single head. Let $W_Q, W_K, W_V \in \mathbb{R}^{d \times D}$ denote the weight matrices of Q, K, and V projection matrices, respectively. The output from the Transformer model is formulated as:

$$A_i = \frac{X_i W_Q W_K^T X_i^T}{\sqrt{D}} \tag{9}$$

$$S_i = \text{Softmax}(A_i) = \text{Softmax}\left(\frac{X_i W_Q W_K^T X_i^T}{\sqrt{D}}\right) \tag{10}$$

$$F_i = S_i X_i W_V \tag{11}$$

where D is the hidden dimension of the model.

We train the above Transformer using the Frobenius norm squared loss function.

$$\mathcal{L}_i = \frac{\|F_i - Y_i\|_F^2}{ND} \tag{12}$$

Note that, for brevity, we use S_i instead of $A_i^l(\phi)$ in Equation 4 to denote the attention matrix. In addition, we write the cross-modal attention matrix $\chi_i(\phi)$ as χ_i.

Assumptions. We assume that the model embedding X_i and weight matrices $W_{\{Q,K,V\}}$ have bounded norms. Let X, Q, K, and V be their corresponding upper bounds. In addition, the attention matrix S_i also has a bounded norm S as each entity is bounded between 0 and 1. Furthermore, we assume that the model always makes a reasonable prediction which is not far from the ground-truth label. In other words, the loss is bounded and we denote the upper bound of the norm of the difference between the predicted output and ground-truth label as $\alpha = \sup_i \|F_i - y_i\|$.

Bound of target alignment trajectories. From the assumption that the alignment trajectories of proxy and target models are close, we have:

$$\|\chi_i^{\text{proxy}} - \chi_i^{\text{target}}\| \leq T, \quad \forall i. \tag{13}$$

For i and j in the same cluster C_k based on the proxy model, we have:

$$\|\chi_i^{\text{proxy}} - \chi_j^{\text{proxy}}\| \leq K_{ij}. \tag{14}$$

Using the triangle inequality:

$$\begin{aligned} \|\chi_i^{\text{target}} - \chi_j^{\text{target}}\| &\leq \|\chi_i^{\text{target}} - \chi_i^{\text{proxy}}\| + \|\chi_i^{\text{proxy}} - \chi_j^{\text{proxy}}\| + \|\chi_j^{\text{proxy}} - \chi_j^{\text{target}}\| \\ &\leq 2T + K_{ij} = \epsilon'. \end{aligned} \tag{15}$$

Therefore, for two samples i and j in the same cluster C_k at any iteration t, we have:

$$\|\chi_i^{\text{target}} - \chi_j^{\text{target}}\| \leq \epsilon', \quad \forall t. \tag{16}$$

Gradient decomposition. Before bounding the gradient difference of the target model, we introduce the expressions of the model gradient w.r.t different weight matrices in the following lemma.

Lemma 1. *Jacobians of the attention weight matrices $W_{\{Q,K,V\}}$ have the following form:*

1. $\nabla_{W_V}\mathcal{L}_i(\phi^t) = \frac{2}{ND}(F_i - y_i)(S_iX_i \otimes I_D) = \frac{2}{ND}(S_iX_iW_V - y_i)(S_iX_i \otimes I_D)$
2. $\nabla_{W_Q}\mathcal{L}_i(\phi^t) = \frac{2}{ND}(S_iX_iW_V - y_i)(I_N \otimes W_V^T X_i^T)\frac{\partial S_i}{\partial A_i}\frac{X_i \otimes X_i W_K}{\sqrt{D}}$
3. $\nabla_{W_K}\mathcal{N}_i(\phi^t) = \frac{2}{ND}(S_iX_iW_V - y_i)(I_N \otimes W_V^T X_i^T)\frac{\partial S_i}{\partial A_i}(\frac{X_i \otimes X_i W_Q}{\sqrt{D}})\Lambda_{D,d}$

where I_L denotes the identity matrix of size L and $\Lambda_{D,d}$ is the commutation matrix.

Lemma 2. *Bound on the Frobenius norm of the Jacobian of the attention matrix:*

$$\|\nabla_{A_i} S_i\|_F \leq \frac{\sqrt{N}}{2} \tag{17}$$

A.1 Proof of Theorem 4.1

Proof. Since we have considered a simplified model with only three parameters, W_Q, W_K and W_V, our goal now is to find $\|\nabla\mathcal{L}_i(\phi^t_{target}) - \nabla\mathcal{L}_j(\phi^t_{target})\|_F$ which can be expanded as

$$\|\nabla\mathcal{L}_i(\phi^t_{\text{target}}) - \nabla\mathcal{L}_j(\phi^t_{\text{target}})\|_F = \sqrt{\begin{array}{l} \|\nabla_{W_V}\mathcal{L}_i(\phi^t_{\text{target}}) - \nabla_{W_V}\mathcal{L}_j(\phi^t_{\text{target}})\|_F^2 \\ + \|\nabla_{W_Q}\mathcal{L}_i(\phi^t_{\text{target}}) - \nabla_{W_Q}\mathcal{L}_j(\phi^t_{\text{target}})\|_F^2 \\ + \|\nabla_{W_K}\mathcal{L}_i(\phi^t_{\text{target}}) - \nabla_{W_K}\mathcal{L}_j(\phi^t_{\text{target}})\|_F^2 \end{array}} \tag{18}$$

To bound the LHS in Equation 18, we first bound each term in the RHS.

Bound for W_V. We simplify the first term in Equation 18 as follows:

$$\begin{aligned}
&\|\nabla_{W_V}\mathcal{L}_i(\phi^t_{\text{target}}) - \nabla_{W_V}\mathcal{L}_j(\phi^t_{\text{target}})\|_F \\
&\overset{(i)}{=} \frac{2}{ND}\|(S_iX_iW_V - y_i)(S_iX_i \otimes I_D) - (S_jX_jW_V - y_j)(S_jX_j \otimes I_D)\|_F \\
&= \frac{2}{ND}\|S_iX_iW_V((S_iX_i - S_jX_j) \otimes I_D) + (S_iX_i - S_jX_j)W_V(S_jX_j \otimes I_D) \\
&\quad + y_j((S_jX_j - S_iX_i) \otimes I_D) + (y_j - y_i)(S_iX_i \otimes I_D)\|_F \\
&\overset{(ii)}{\leq} \frac{2}{ND}\|S_iX_iW_V((S_iX_i - S_jX_j) \otimes I_D)\|_F + \frac{2}{ND}\|(S_iX_i - S_jX_j)W_V(S_jX_j \otimes I_D)\|_F \\
&\quad + \frac{2}{ND}\|y_j((S_jX_j - S_iX_i) \otimes I_D)\|_F + \frac{2}{ND}\|(y_j - y_i)(S_iX_i \otimes I_D)\|_F \\
&\overset{(iii)}{\leq} \frac{2SXV}{N}\|(S_i(X_i - X_j) + (S_i - S_j)X_j\|_F + \frac{2SXV}{N}\|(S_i(X_i - X_j) + (S_i - S_j)X_j\|_F \\
&\quad + \frac{2}{N}B\|(S_i(X_i - X_j) + (S_i - S_j)X_j\|_F + \frac{4BSX}{N} \\
&\overset{(iv)}{\leq} \frac{8S^2X^2V}{N} + \frac{4SX^2V}{N}\|S_i - S_j\|_F + \frac{4BSX}{N} + \frac{2BX}{N}\|S_i - S_j\|_F + \frac{4BSX}{N} \\
&\overset{(v)}{\leq} \left(\frac{4SX^2V}{N} + \frac{2BX}{N}\right)\|S_i - S_j\|_F + \frac{8SX}{N}(B + SXV)
\end{aligned} \tag{19}$$

where (i) comes from Lemma 1; (ii) uses triangle inequality; (iii) applies assumed bounds; (iv) uses $\|X_i - X_j\|_F \leq 2X$; (v) rearranging.

Now, consider the following relation for the attention matrix:

$$S_i = \chi_i + \chi_i^T + S_i^r \tag{20}$$

where χ_i is the cross-modal attention part (bottom-left section) of the attention matrix that is considered in the actual experiments. χ_i is the same shape as χ_i with rest of the entries being zero. Similary, S_i^r is the same shape as S_i, with entries in the cross-modal attention part being zero and all other entries being non-zero. Now, consider the following difference:

$$\begin{aligned}
&\|\nabla_{W_V}\mathcal{N}_i(\phi^t_{\text{target}}) - \nabla_{W_V}\mathcal{N}_j(\phi^t_{\text{target}})\|_F \\
&\overset{(i)}{\leq} \left(\frac{4SX^2V}{N} + \frac{2BX}{N}\right)\|\chi_i - \chi_j + \chi_i^T - \chi_j^T + S_i^r - S_j^r\|_F + \frac{8SX}{N}(B + SXV) \\
&\overset{(ii)}{\leq} \left(\frac{8SX^2V}{N} + \frac{4BX}{N}\right)\|\chi_i - \chi_j\|_F + \left(\frac{4SX^2V}{N} + \frac{2BX}{N}\right)\|S_i^r - S_j^r\|_F + \frac{8SX}{N}(B + SXV) \\
&\overset{(iii)}{\leq} \left(\frac{8SX^2V}{N} + \frac{4BX}{N}\right)\|\chi_i - \chi_j\|_F + \frac{4SX}{L}(4SXV + 3B) \\
&= \left(\frac{8SX^2V}{N} + \frac{4BX}{N}\right)\epsilon' + \frac{4SX}{N}(4SXV + 3B)
\end{aligned} \tag{21}$$

where (i) from equation 20; (ii) applies triangle inequality; (iii) $\|S_i^r - S_j^r\|_F \leq 2S$.

Let $f_1(\epsilon') = \left(\frac{8SX^2V}{N} + \frac{4BX}{N}\right)\epsilon' + \frac{4SX}{N}(4SXV + 3B)$ be an affine function of ϵ', we can bound the gradient difference w.r.t W_V as

$$\|\nabla_{W_V}\mathcal{L}_i(\phi^t_{\text{target}}) - \nabla_{W_V}\mathcal{L}_j(\phi^t_{\text{target}})\|_F \leq f_1(\epsilon') \tag{22}$$

Bound for W_Q. We simplify the second term in Equation 18 as follows:

$$\begin{aligned}
&\|\nabla_{W_Q}\mathcal{L}_i(\phi^t_{\text{target}}) - \nabla_{W_Q}\mathcal{L}_j(\phi^t_{\text{target}})\|_F^2 \\
&\overset{(i)}{=} \frac{2}{ND}\left\|(S_iX_iW_V - y_i)(I_N \otimes W_V^T X_i^T)\frac{\partial S_i}{\partial A_i}\frac{X_i \otimes X_iW_K}{\sqrt{D}} - (S_jX_jW_V - y_j)(I_N \otimes W_V^T X_j^T)\frac{\partial S_j}{\partial A_j}\frac{X_j \otimes X_jW_K}{\sqrt{D}}\right\|_F \\
&\overset{(ii)}{=} \frac{2}{ND}\bigg\|(S_iX_iW_V - y_i)\,(I_N \otimes W_V^T X_i^T)\,\frac{\partial S_i}{\partial A_i}\,\Big(\frac{X_i \otimes X_iW_K - X_j \otimes X_jW_K}{\sqrt{D}}\Big) \\
&\quad + \Big((S_iX_iW_V - y_i)\,(I_N \otimes W_V^T X_i^T)\,\frac{\partial S_i}{\partial A_i} \;-\; (S_jX_jW_V - y_j)\,(I_N \otimes W_V^T X_j^T)\,\frac{\partial S_j}{\partial A_j}\Big)\;\frac{X_j \otimes X_j\,W_K}{\sqrt{D}}\bigg\|_F \\
&= \frac{2}{ND}\bigg\|(S_iX_iW_V - y_i)\,(I_N \otimes W_V^T X_i^T)\,\frac{\partial S_i}{\partial A_i}\,\Big(\frac{X_i \otimes (X_i - X_j)W_K + (X_i - X_j) \otimes X_jW_K}{\sqrt{D}}\Big) \\
&\quad + \bigg[(S_iX_iW_V - y_i)\,\Big((I_N \otimes W_V^T X_i^T)\Big(\frac{\partial S_i}{\partial A_i} - \frac{\partial S_j}{\partial A_j}\Big) + (I_N \otimes W_V^T(X_i^T - X_j^T))\,\frac{\partial S_j}{\partial A_j}\Big) \\
&\qquad + \bigg\{\Big(S_i(X_i - X_j) + (S_i - S_j)X_j\Big)W_V + (y_j - y_i)\bigg\}(I_N \otimes W_V^T X_j^T)\,\frac{\partial S_j}{\partial A_j}\,\bigg]\,\frac{X_j \otimes X_j\,W_K}{\sqrt{D}}\bigg\|_F \\
&\overset{(iii)}{\leq} \frac{8\sqrt{N}VX^3K}{D^{3/2}}\|F_i - y_i\|_F + \frac{2\sqrt{N}SX^4V^2K}{D^{3/2}} + \frac{\sqrt{N}V^2X^4K}{D^{3/2}}\|S_i - S_j\|_F + \frac{2B\sqrt{N}VX^3K}{D^{3/2}} \\
&= \frac{\sqrt{N}V^2X^4K}{D^{3/2}}\|S_i - S_j\|_F + \frac{2\sqrt{N}VX^3K}{D^{3/2}}\Big[4\alpha + SVX + B\Big]
\end{aligned} \tag{23}$$

where (i) from Lemma 1; (ii) rearranging; (iii) uses triangle inequality and applies the assumptions on the bounds.

Plugging Equation 20 into Equation 23, we have

$$\begin{aligned}\|\nabla_{W_Q}\mathcal{L}_i(\phi^t_{\text{target}}) - \nabla_{W_Q}\mathcal{L}_j(\phi^t_{\text{target}})\|_F &\leq \frac{2\sqrt{N}V^2X^4K}{D^{3/2}}\|\chi_i - \chi_j\|_F + \frac{2\sqrt{N}VX^3K}{D^{3/2}}\Big[4\alpha + 2SVX + B\Big] \\ &= \frac{2\sqrt{N}V^2X^4K}{D^{3/2}} \cdot \epsilon' + \frac{2\sqrt{N}VX^3K}{D^{3/2}}\Big[4\alpha + 2SVX + B\Big] \qquad (24)\end{aligned}$$

Let $f_2(\epsilon') = \frac{2\sqrt{N}V^2X^4K}{D^{3/2}} \cdot \epsilon' + \frac{2\sqrt{N}VX^3K}{D^{3/2}}\Big[4\alpha + 2SVX + B\Big]$ be an affine function of ϵ', we can bound the gradient difference w.r.t W_Q as

$$\|\nabla_{W_Q}\mathcal{L}_i(\phi^t_{\text{target}}) - \nabla_{W_Q}\mathcal{L}_j(\phi^t_{\text{target}})\|_F \leq f_2(\epsilon') \qquad (25)$$

Bound for W_K. Similarly, we have the following bound for the third term

$$\|\nabla_{W_K}\mathcal{L}_i(\phi^t_{\text{target}}) - \nabla_{W_K}\mathcal{L}_j(\phi^t_{\text{target}})\|_F \leq f_3(\epsilon') \qquad (26)$$

where $f_3(\epsilon') = \frac{2\sqrt{N}V^2X^4Q}{D^{3/2}} \cdot \epsilon' + \frac{2\sqrt{N}VX^3Q}{D^{3/2}}\Big[4\alpha + 2SVX + B\Big]$ be affine function of ϵ'.

Plugging all these expressions in 18, we have the following expression for the upper bound of the target-model gradient difference

$$\|\nabla\mathcal{L}_i(\phi^t_{\text{target}}) - \nabla\mathcal{L}_j(\phi^t_{\text{target}})\|_F \leq \sqrt{f_1(\epsilon')^2 + f_2(\epsilon')^2 + f_3(\epsilon')^2} \qquad (27)$$

Using Cauchy-Schwarz,

$$\begin{aligned}\|\nabla\mathcal{L}_i(\phi^t_{\text{target}}) - \nabla\mathcal{L}_j(\phi^t_{\text{target}})\|_F &\leq \sqrt{f_1(\epsilon')^2 + f_2(\epsilon')^2 + f_3(\epsilon')^2} \\ &\leq |f_1(\epsilon')| + |f_2(\epsilon')| + |f_3(\epsilon')| \\ &= \underbrace{\left[\frac{2\sqrt{N}V^2X^4(K+Q)}{D^{3/2}} + \left(\frac{8SX^2V}{N} + \frac{4BX}{N}\right)\right]}_{c_1} \cdot \epsilon' \\ &\quad + \underbrace{\frac{2\sqrt{N}VX^3}{D^{3/2}}\Big[(4\alpha + 2SVX + B)(Q+K)\Big] + \frac{4SX}{N}(4SXV + 3B)}_{c_2}\end{aligned}$$

Assume the model weights Q, K, and V are bounded by $\frac{c}{\sqrt{3}}$, so that $\|\phi^t_{\text{target}}\| \leq c$. We further assume the ground-truth embedding B and loss α are bounded. Such boundedness assumptions are standard in theoretical analyses of Transformers, including generalization and stability (Li et al., 2023b), as well as gradient-based optimization dynamics (Song et al., 2024). In practice, they are enforced by weight decay, and layer/activation normalization (Xiong et al., 2020).

Since X passes through an RMS normalization layer with gain g, we have $X \leq g\sqrt{ND}$. Under this setting, the constants c_1 and c_2 simplify to:

$$c_1 = \mathcal{O}\Big(\tfrac{4}{\sqrt{3}} N^2 \sqrt{ND}\, c^3 g^4\Big), \tag{28}$$

$$c_2 = \mathcal{O}\Big(\tfrac{8}{3\sqrt{3}} N^3 \sqrt{D}\, c^2 g^4\Big). \tag{29}$$

If the RMS gain is sufficiently small, i.e.,

$$g < N^{-5/8} D^{-1/8} c^{-3/4} \quad \Leftrightarrow \quad g^4 < N^{-5/2} D^{-1/2} c^{-3},$$

then the bounds on c_1 and c_2 reduce to:

$$c_1 \leq \tfrac{4}{\sqrt{3}}, \tag{30}$$

$$c_2 \leq \tfrac{8\sqrt{N}}{3\sqrt{3}c}. \tag{31}$$

Therefore, the gradient distance can be bounded as

$$\|\nabla \mathcal{L}_i(\phi^t_{\text{target}}) - \nabla \mathcal{L}_j(\phi^t_{\text{target}})\|_F \leq \tfrac{4}{\sqrt{3}} \cdot \epsilon' + \tfrac{8\sqrt{N}}{3\sqrt{3}c}.$$

□

A.2 Proof of Theorem 4.2

We first have the following lemma.

Lemma 3 ((Zhang, 2023), Theorem 1.8)**.** *Let f: $\mathbb{R}^n \rightarrow \mathbb{R}$ be a twice continuously differentiable function. Given $x, y \in \mathbb{R}^n$, for every $z = x + t(y - x)$ where $t \in [0, 1]$, there exists a constant, $\zeta = \sup \|\nabla^2 f(x)\|$, such that*

$$\|\nabla f(z) - \nabla f(x)\|_F \leq \zeta \|z - x\|_F \tag{32}$$

Proof.

$$\|\nabla \mathcal{L}_i(\phi^{t_z}_{\text{target}}) - \nabla \mathcal{L}_j(\phi^{t_z}_{\text{target}})\|_F \tag{33}$$

$$\overset{(i)}{\leq} \|\nabla \mathcal{L}_i(\phi^{t_z}_{\text{target}}) - \nabla \mathcal{L}_i(\phi^{t_1}_{\text{target}})\|_F + \|\nabla \mathcal{L}_j(\phi^{t_z}_{\text{target}}) - \nabla \mathcal{L}_j(\phi^{t_1}_{\text{target}})\|_F + \|\nabla \mathcal{L}_i(\phi^{t_1}_{\text{target}}) - \nabla \mathcal{L}_j(\phi^{t_1}_{\text{target}})\|_F \tag{34}$$

$$\overset{(ii)}{\leq} 2\beta \|\phi^{t_z}_{\text{target}} - \phi^{t_1}_{\text{target}}\| + \Delta^{t_1}_{ij} \tag{35}$$

$$\overset{(iii)}{\leq} 2\delta\beta + \Delta^{t_1}_{ij} \tag{36}$$

where (i) from triangle inequality; (ii) from Lemma 3; (iii) from bounded curvature assumption and the definition of δ. Similarly, we have

$$\|\nabla \mathcal{L}_i(\phi^{t_z}_{\text{target}}) - \nabla \mathcal{L}_j(\phi^{t_z}_{\text{target}})\|_F \leq 2\delta\beta + \Delta^{t_2}_{ij} \tag{37}$$

Combining these two above inequalities, we have

$$\|\nabla \mathcal{L}_i(\phi^{t_z}_{\text{target}}) - \nabla \mathcal{L}_j(\phi^{t_z}_{\text{target}})\|_F \leq 2\delta\beta + \max\{\Delta^{t_1}_{ij}, \Delta^{t_2}_{ij}\} \tag{38}$$

□

A.3 CONVERGENCE OF XMAS

Corollary A.1 (Convergence of XMAS)**.** *Under the assumptions of Theorem 4.2, applying incremental gradient methods with stepsize η on subsets found by XMAS, converges to a neighborhood of the solution ϕ^* found by training on full data:*

$$\|\phi^{t+1} - \phi^*\|^2 \leq (1 - \eta c')^{t+1} \|\phi^t - \phi^*\|^2 + \frac{2\xi R'}{c'^2} + \eta B^2 \left(\frac{r_{\min}}{k}\right)^2 g_{\max}^2 \tag{39}$$

where $c' \leq \|H\|$, B is the total size of the subset, $g_{\max}$ is the largest gradient norm of individual examples during training, $r_{\min}, r_{\max}$, are the size of the smallest and largest clusters, $R' = \min\{d_0, Bg_{\max} + \xi/c'\}$ and $d_0 = \|\phi^0 - \phi^*\|$ is the initial distance to the optimal solution ϕ^*, and $\xi = K[r_{\min}\Delta_2 + (r_{\max} - r_{\min})g_{\max}]$.

Our proof is similar to the one in S2L (Yang et al., 2024, Appendix A.2). We provide the details in order to ensure completeness.

Proof. Without loss of generality, assume we select k examples from each cluster and we have $k \leq \min_{j\in[K]} |C_j|$. Then the error of the subset in capturing the full gradient will be

$$\xi \leq \sum_j (|C_j| - k) \left(\bar{g}_j + \Delta\right), \tag{40}$$

where $\bar{g}_j$ is the norm of the average gradient of all samples from C_j. Here Δ is the maximum error in gradients between two different samples.

In practice, we can weight elements of the subset by $r_{\min}/k$, which has a similar effect to scaling the step size when training on the subset. Let $g_{\max} = \max_j \|g_j\|$ be the maximum gradient norm during training, $r_{\max} = \max_j |C_j|$, $r_{\min} = \min_j |C_j|$. Then, we get

$$\xi' \leq \sum_j \left[(r_{\min} - k)\Delta + (|C_j| - r_{\min})(\bar{g}_j + \Delta)\right] \tag{41}$$

$$\leq K \left[r_{\min}\Delta + (r_{\max} - r_{\min})g_{\max}\right]. \tag{42}$$

The first term in the RHS of Eq. (7) is the error of the subset selected from C_j to capture its full gradient and the second term is due to selecting the same number of examples, k, from the larger clusters. Using the above error and following the proof of Theorem 1 in Mirzasoleiman et al. (2020), for a constant step size $\alpha \leq 1/c$ we get:

$$\|\theta^{t+1} - \theta^*\|^2 \leq (1 - \alpha c)^{t+1} \|\theta^t - \theta^*\|^2 + \frac{2\xi' R}{c^2} + \alpha B^2 \left(\frac{r_{\min}}{k}\right)^2 g_{\max}^2, \tag{43}$$

where $c \leq \|H\|$, and $B = k \cdot K$ is the total size of the subset, $R = \min\{d_0, Bg_{\max} + \xi'/c\}$ and $d_0 = \|\theta^0 - \theta^*\|$ is the initial distance to the optimal solution θ^*.

If $k \geq |C_j|$ for any cluster C_j, one can simply add $(r_{\min}/k - 1) \cdot \hat{g}_j$ to ξ' for the corresponding clusters, where $\hat{g}_j$ is the norm of the total gradient of cluster C_j and we replace $r_{\min}$ in Eq. (7) with the size of smallest cluster that has larger than k examples. □

A.4 Proofs of Lemmas

Proof of Lemma 1.

Proof.

1. Applying chain rule and simplifying, we get

$$\begin{aligned}\nabla_{W_V}\mathcal{L}_i(\phi^t) &= \nabla_{F_i}\mathcal{L}_i(\phi^t)\cdot\nabla_{W_V}F_i(\phi^t)\\ &= (F_i - Y_i)(S_iX_i \otimes I_D)\end{aligned}$$

2. Applying chain rule and simplifying, we get

$$\begin{aligned}\nabla_{W_Q}\mathcal{L}_i(\phi^t) &= \nabla_{F_i}\mathcal{L}_i(\phi^t)\cdot\nabla_{S_i}F_i(\phi^t)\cdot\nabla_{A_i}S_i(\phi^t)\cdot\nabla_{W_Q}A_i(\phi^t)\\ &= (F_i - Y_i)(I_N \otimes W_V^T X_i^T)\cdot\nabla_{A_i}S_i(\phi^t)\cdot(\frac{X_i \otimes X_iW_K}{\sqrt{D}})\end{aligned}$$

3. Proceeding as above for $\nabla_{W_Q}\mathcal{L}_i(\phi^t)$, we get

$$\nabla_{W_K}\mathcal{L}_i(\phi^t) = (S_iX_iW_V - y_i)(I_N \otimes W_V^T X_i^T)\frac{\partial S_i}{\partial A_i}\left(\frac{X_i \otimes X_iW_Q}{\sqrt{D}}\right)\Lambda_{d,D}$$

where $\Lambda_{d,D}$ is the commutation matrix.

□

Proof of Lemma 2.

Proof. For a single sample, i, we have

$$S_i = softmax(A_i)$$

where softmax is applied row-wise to A. The expression for a single element in S_i present at (j, k) is given as:

$$(S_i)_{jk} = \frac{e^{(A_i)_{jk}}}{\sum_{z=1}^{L} e^{(A_i)_{jz}}}$$

Now, clearly, we have

$$\frac{\partial(S_i)_{jk}}{\partial(A_i)_{mn}} = 0 \ \text{ if } j \neq m$$

Now for a specific row, i, the Jacobian of the i-th row of S_i wrt the i-th row of A_i is given as

$$\begin{aligned}J_{kn}^{(j)} = \frac{\partial(S_i)_{jk}}{\partial(A_i)_{jn}} &= \begin{cases}(S_i)_{jk}\Big(1-(S_i)_{jk}\Big) & \text{if } k = n\\ -(S_i)_{jk}(S_i)_{jn} & \text{if } k \neq n\end{cases}\\ &= (S_i)_{jk}\Big(\delta_{kn} - (S_i)_{kn}\Big)\end{aligned}$$

where δ_{jm} is the Kronecker delta (1 if j=m, 0 otherwise).

Now the complete Jacobian of the attention matrix can be written as follows:

$$(\nabla_{A_i}S_i)_{jkmn} = \delta_{jm}(S_i)_{jk}\Big(\delta_{kn} - (S_i)_{kn}\Big)$$

where δ_{jm} and δ_{kn} are the Kronecker deltas.

Following is the expression for the Frobenius Norm of this Jacobian:

$$
\begin{aligned}
\|(\nabla_{A_i} S_i)_{jkmn}\|_F^2 &= \sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N}\left(\frac{\partial(S_i)_{jk}}{\partial(A_i)_{mn}}\right)^2 \\
&= \sum_{j=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N}\left(\frac{\partial(S_i)_{jk}}{\partial(A_i)_{jn}}\right)^2 \\
\|(\nabla_{A_i} S_i)_{jkmn}\|_F^2 &= \sum_{j=1}^{N}\|J^{(j)}\|_F^2
\end{aligned}
\tag{44}
$$

Now computing $\|J^{(j)}\|_F$ for a single row j. Also, let $\xi_p^{(j)} = \sum_{q=1}^{N}(S_i)_{jq}^p$ Then we have

$$
\begin{aligned}
\|J^{(j)}\|_F^2 &= \sum_{k=1}^{N}\sum_{n=1}^{N}\left(J_{kn}^{(j)}\right) \\
&= \sum_{k=1}^{N}\Big((S_i)_{jk}(1-(S_i)_{jk})\Big)^2 + \sum_{k=1}^{N}\sum_{n\neq k}\Big(-(S_i)_{jk}(S_i)_{jn}\Big)^2 \\
&= \xi_2^{(j)} - 2\xi_3^{(j)} + \left(\xi_2^{(j)}\right)^2
\end{aligned}
$$

Now the maximum value of $\|J^{(j)}\|_F^2$ under constraints, $\xi_1^{(j)} = 1$ and $(S_i)_{jq} \geq 0 \;\; \forall q \in \{0, 1, 2, \ldots, N\}$, is $\frac{1}{4}$ and is achieved when exactly two q, $(S_i)_{jq} = \frac{1}{2}$ when $N \geq 2$. Plugging this in equation 44, we get

$$
\|(\nabla_{A_i} S_i)_{jkmn}\|_F^2 = \sum_{j=1}^{N}\|J^{(j)}\|_F^2 \leq \sum_{j=1}^{N}\left(\frac{1}{4}\right) = \frac{N}{4}
$$

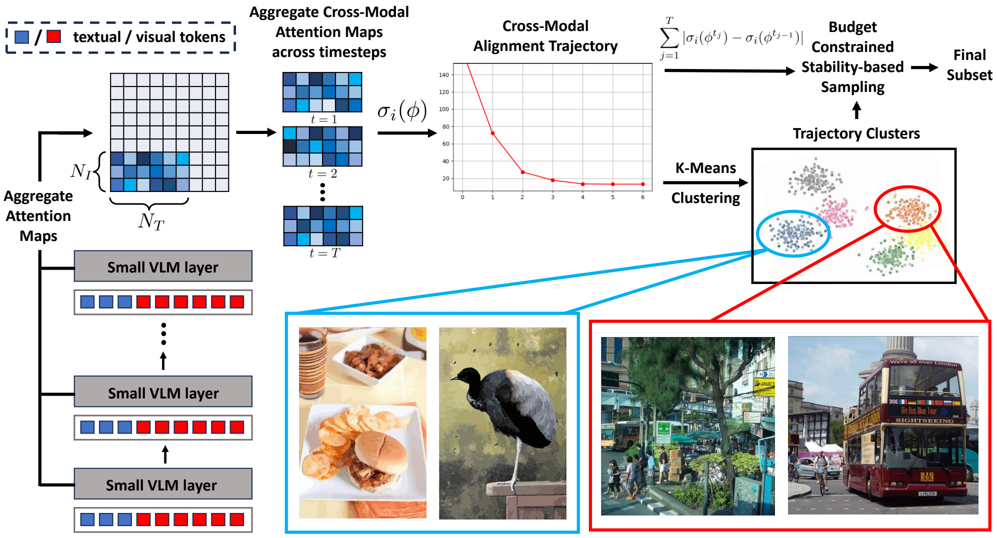
Therefore, we have

$$
\|(\nabla_{A_i} S_i)_{jkmn}\|_F \leq \frac{\sqrt{N}}{2}
$$

□

B Additional Experimental Settings

Models. For the target LVLMs, we use the pre-trained LLaVA-1.5-7B, LLaVA-1.5-13B models Liu et al. (2024a), LLaVA-1.6-Mistral-7B Li et al. (2024) and Phi-3.5-Vision-Instruct‘Abdin et al. (2024). For the proxy models, we use the TinyLLaVA (Zhou et al., 2024) with 2 different scales 0.5B and 2.0B. The default one is TinyLLaVA 2.0B due to its superior performance. Table 2 summarizes the language model and vision encoder of different models used in our experiment.



■ / ■ textual / visual tokens
Aggregate Cross-Modal Attention Maps across timesteps
$t = 1$
$t = 2$
$t = T$
N_I
N_T
Aggregate Attention Maps
Small VLM layer
Small VLM layer
Small VLM layer
$\sigma_i(\phi)$
Cross-Modal Alignment Trajectory
140
120
100
80
60
40
20
0
1
2
3
4
5
6
$\sum_{j=1}^{T} |\sigma_i(\phi^{t_j}) - \sigma_i(\phi^{t_{j-1}})|$
Budget Constrained Stability-based Sampling
Final Subset
Trajectory Clusters
K-Means Clustering

Figure 5: XMAS employs a small proxy VLM to find alignment trajectory for examples in the fine-tuning data. Examples with similar alignment trajectory have similar gradients during instruction tuning. Then, it clusters the alignment trajectories and sample a balanced subset of examples with more stable trajectories from the clusters.

Table 2: Details of the model architectures used in our experiments are provided below. Model names correspond to their repository names on HuggingFace.

Model	Language model	Vision encoder
LLaVA-1.5-13B	lmsys/vicuna-13b-v1.5	openai/clip-vit-large-patch14
LLaVA-1.5-7B	lmsys/vicuna-7b-v1.5	openai/clip-vit-large-patch14
TinyLLaVA 2.0B	stablelm/stablelm-2-zerphyr-1_6b	openai/clip-vit-large-patch14
TinyLLaVA 0.5B	Qwen/Qwen2-0.5B	google/siglip-so400m-patch14-384

VIT datasets. We apply coreset selection to two distinct vision instruction tuning (VIT) datasets: LLaVA 665k Liu et al. (2024a) and Vision-Flan Xu et al. (2024), both of which are widely used benchmarks for evaluating multimodal instruction-following models. LLaVA 665k comprises approximately 665,000 VIT examples aggregated from 12 diverse vision-language datasets. These datasets span a wide range of tasks, such as image captioning, visual question answering, and visual reasoning, providing a comprehensive training mixture for instruction-tuned large vision-language models. The examples are automatically aligned and instruction-formatted, making the dataset suitable for large-scale fine-tuning. In contrast, Vision-Flan is a more task-structured benchmark composed of 191 individual vision-language tasks. Each task includes around 1,000 high-quality, expert-labeled VIT examples, leading to a total of about 186,000 samples. Unlike LLaVA 665k, which merges data across tasks, Vision-Flan preserves a task-level granularity, allowing for more fine-grained evaluation and task-specific data selection strategies.

Training details. In all experiments, we fine-tune the target models using LoRA Hu et al. (2022) for one epoch regardless of the subset size. We strictly follow the official finetuning hyperparameters specified in LLaVA-1.5. For Phi-3.5-Vision-Instruct, we used the settings in `https://github.com/microsoft/PhiCookBook`. For proxy models, we train full model (i.e., without LoRA) for one epoch, following the official hyperparameters specified in TinyLLaVA. This results in a total of $T = 7$ checkpoints for the

Table 3: A comparative analysis of the average relative performance of LLaVA-1.5-7B using different methods at different sampling ratios on the LLaVA 665k and Vision-Flan datasets.

Method	LLaVA-665K				Vision-Flan		
	10%	20%	30%	50%	5%	10%	15%
Rand	93.5	96.3	97.4	99.2	92.7	96.2	98.4
MP (Marion et al., 2023)	92.4	95.4	97.9	99.6	92.9	95.6	97.7
HL (Zhou et al., 2023b)	90.2	95.4	96.6	99.8	91.4	95.8	97.5
EL2N (Paul et al., 2021)	93.3	95.0	95.6	99.3	88.7	91.9	87.4
D2 Pruning (Maharana et al., 2023)	91.8	94.1	95.5	97.7	92.4	97.6	99.0
SemDeDup (Abbas et al., 2023)	91.4	95.9	95.4	96.8	92.6	93.0	96.3
CLIP-Score (Hessel et al., 2021)	86.1	90.8	92.3	96.2	93.1	96.6	98.1
Self-Sup (Sorscher et al., 2022)	85.9	91.6	94.6	97.8	85.7	92.4	99.1
Self-Filter (Chen et al., 2024)	89.3	88.4	93.7	97.7	90.3	93.6	94.5
COINCIDE (Lee et al., 2024)	94.1	96.6	98.7	98.8	92.5	97.0	98.3
XMAS (**Ours**)	**95.4**	**97.1**	**99.2**	**100.0**	**96.1**	**98.9**	**100.2**

trajectory. For distributed training, we use DeepSpeed Rasley et al. (2020). For K-means, we set the number of clusters K to 1000 and use the GPU version of the faiss library Johnson et al. (2019).

Evaluation datasets. We evaluate the performance of the fine-tuned target models across four core capabilities: reasoning, hallucination resistance, visual perception, and cognition. To comprehensively assess these abilities, we utilize a diverse suite of both academic-task-oriented benchmarks and recent benchmarks tailored for instruction-following LMMs, totaling ten evaluation datasets.

For visual perception and cognition, we include VQAv2Goyal et al. (2017) and GQAHudson & Manning (2019), which require open-ended answers to visual questions, assessing the model's ability to understand and interpret images. VizWizGurari et al. (2018), a dataset comprising real-world images taken by visually impaired users, is used to test the model's zero-shot generalization capabilities in a more challenging, accessibility-focused setting. TextVQASingh et al. (2019) measures performance on text-rich visual inputs, challenging models to combine OCR with multimodal reasoning. For science-focused reasoning, we adopt the image subset of ScienceQA Lu et al. (2022), which consists of multiple-choice scientific questions accompanied by relevant visual content.

To evaluate hallucination behavior, we employ POPE Li et al. (2023a), which measures the model's tendency to generate factually incorrect information in multimodal contexts. POPE includes three subsets—random, common, and adversarial samples from the COCO dataset and we report the average F1 score across all splits.

For general reasoning and robustness, we use several recently proposed benchmarks. MME-PerceptionLiang et al. (2024) tests perception using binary (yes/no) questions based on visual content, while MMBenchLiu et al. (2024b) evaluates the robustness of multiple-choice answers across a broad range of tasks. MMVetYu et al. (2023) and LLaVABenchLiu et al. (2023) focus on visual conversation abilities, evaluating both the correctness and helpfulness of model responses using GPT-4 as a judge.

By covering a wide spectrum of domains, ranging from scientific reasoning and accessibility to free-form visual conversations, this evaluation protocol provides a comprehensive measure of how well the fine-tuned models generalize across real-world and task-specific scenarios.

Evaluation metric. We follow the same evaluation protocols outlined in LLaVA-1.5. Similar to COINCIDE, we measure the relative performance as (model performance / full-finetuned performance) × 100% to assess

Table 4: Average relative performances (ARP) over full data when training LLaVa-1.5-7B on 10% subsets of LLaVA 665K found by XMAS when using different (a) alignment score, (b) matrix aggegation, and (c) instability score.

Score	ARP
Sum values	**95.4**
Concat values	93.0
Sing vector	93.4

(a) Alignment score

Method	ARP
Sum	**95.4**
Concat text	92.9
Concat vision	92.1

(b) Matrix aggregation

Score	ARP
Sum abs diff	**95.4**
Sum sq diff	94.8
Variance	93.9

(c) Instability score

Table 5: Average relative performances (ARP) over full data when training LLaVa-1.5-7B on 10% subsets of LLaVA 665K found by XMAS when using different (a) cluster sampling strategies, (b) attention matrices, and (c) found by COINCIDE with different layer choices.

Attn layer	ARP
First layer	92.9
Last layer	93.7
All layers	**95.4**

(a) Attention layer

Singular values	ARP
All	**95.6**
Top-5	95.4
Top-1	94.2

(b) Singular values

Layer indices	ARP
3,7,11,15,19	**94.1**
4,8,12,16,20	91.6
3,8,13,18,23	92.1

(c) Layer choice

the performance of subsets compared to full dataset. To compare between different methods, we **A**verage the **R**elative **P**erformance (ARP) across all evaluation datasets.

Computational resources. All experiments are conducted using 8 NVIDIA RTX A6000 GPUs.

C Additional experimental results

Quantitative results of LLaVA 665k and Vision-Flan. In this section, we present the detailed Average Relative Performance (ARP) corresponding to the experiments in Section 5.1. The full ARP results underlying Figures 1 are reported in Table 3. As shown, our method consistently achieves the highest performance across various subset budgets on both LLaVA 665k and Vision-Flan datasets. Notably, XMAS is the only approach that consistently outperforms random sampling on LLaVA 665k. On Vision-Flan, XMAS outperforms COINCIDE by nearly 2% when selecting subsets in the 5–15% range.

Choice of alignment score. Table 4a compares different choices of the alignment score (Definition 4.1). For a single checkpoint, instead of summing the top-5 singular values, one option is to concatenate them into a 5-dimensional vector, resulting in an alignment trajectory of size 35 across 7 checkpoints. Another option is to use the singular vector (dimension 576) corresponding to the largest singular value, giving a trajectory of size 4032 in Equation 7. As shown, summing the top-5 singular values outperforms the alternatives by about 2%.

Choice of layer aggregation strategies. In Definition 4.1, we aggregate the cross-modal attention matrices by summing them across all layers, resulting in a single matrix of size $n_T \times n_I$. In this experiment, we benchmark this approach with alternative strategies that stack the attention matrices along either the text or vision dimension, producing matrices of size $Ln_T \times n_I$ or $n_T \times Ln_I$, respectively, where L is the number of layers. As shown in Table 4b, summing the per-layer attention matrices outperforms both stacking approaches. We hypothesize that summation reduces noise and highlights shared structures across layers,

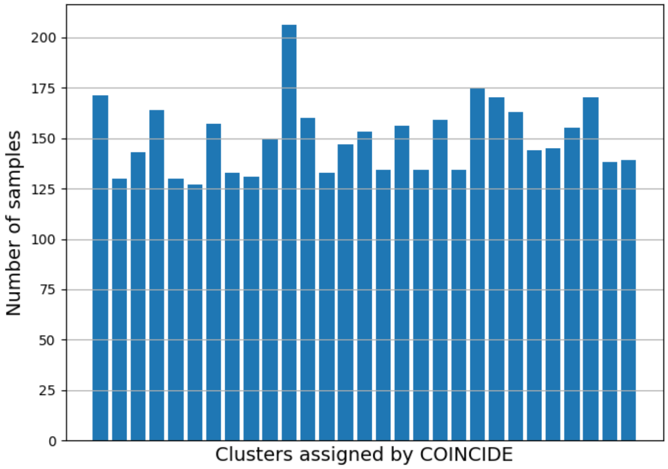


Figure 6: Distribution of concepts, i.e., clusters found by COINCIDE on LLaVA 665k, in the largest cluster found by XMAS. Only concepts with more than 100 samples are shown in the figure.

Table 6: Average relative performances (ARP) comparison of different 10% subsets of the LLaVA 665k dataset across target models. The best method is shown in **bold**, and the second best is <u>underlined</u>. Results marked with * report ARP without VizWiz evaluation due to its unavailability on EvalAI.

Method	**LLaVA-1.5-7B**	**LLaVA-1.5-13B**	**LLaVA-1.6-Mistral-7B***	**Phi-3-Vision-Instruct***
Rand	93.5	93.7	91.9	101.8
MP	92.4	92.5	88.8	101.5
HL	90.2	93.7	78.1	102.2
EL2N	93.3	93.3	87.8	101.7
D2 Pruning	91.8	91.8	89.5	101.7
SemDeDup	91.4	90.8	<u>92.1</u>	<u>102.6</u>
CLIP-Score	86.1	89.9	86.7	102.3
Self-Sup	85.9	90.2	88.9	102.1
Self-Filter	89.3	89.8	88.4	102.4
COINCIDE	<u>94.1</u>	<u>94.6</u>	91.6	101.9
XMAS (**Ours**)	**95.4**	**95.6**	**93.8**	**103.3**

leading to more robust alignment signals. Moreover, this strategy is computationally more efficient, as computing the SVD on the summed matrix is significantly faster than on the higher-dimensional stacked variants. Therefore, we adopt layer-wise summation as our default aggregation method.

Choice of instability score. Table 4c compares alternative definitions of the instability score for alignment trajectories. Our default choice, the sum of absolute differences $S_i = \sum_{j=1}^{T} |\sigma_i(\phi^{t_j}) - \sigma_i(\phi^{t_{j-1}})|$, achieves the best performance (95.4 ARP) as it directly measures cumulative oscillation across checkpoints. Using squared differences $S_i^{\text{sqr}} = \sum_{j=1}^{T} (\sigma_i(\phi^{t_j}) - \sigma_i(\phi^{t_{j-1}}))^2$ slightly reduces performance, since large fluctuations are overweighted. Variance $S_i^{\text{var}} = \frac{1}{T} \sum_{j=1}^{T} (\sigma_i(\phi^{t_j}) - \frac{1}{T} \sum_{k=1}^{T} \sigma_i(\phi^{t_k}))^2$ performs worse, as it only captures the spread of values and ignores the temporal ordering of oscillations.

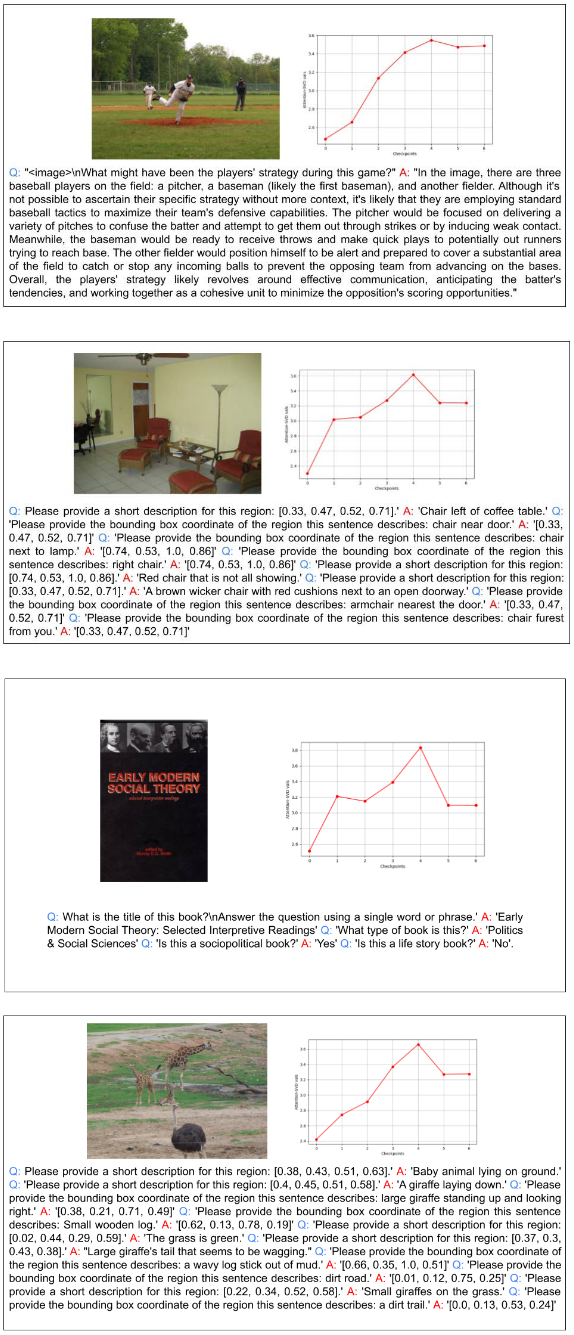
Choice of attention layers. Table 5a illustrates that aggregating cross-modal matrices across all the layers outperforms only using the first or last layer.

Number of singular values. While using all the singular values of the cross-modal attention can capture its full spectrum, it is more computationally more expensive. Indeed, using only top-5 singular values reduces the SVD computation from 51.7 (ms) to 1.77 (ms) on average. Furthermore, using top-5 singular values only harms the average relative performance by 0.2% as detailed in Table 5b while using top-1 singular values decreases the performance more significantly by 1.4%.

COINCIDE is sensitive to the layer choice. We would like to emphasize that COINCIDE's total time includes the cost of training a proxy model and finding layers that encode different concepts, but the cost for "finding layers" are omitted from the time we reported, as we directly used the layers specified in their paper. For finding the layers, COINCIDE needs to try different sets of layers (as much as $\binom{24}{5}$ and retrain the target model on the corresponding subsets. This makes COINCIDE considerably more expensive than our method (although we are not able to calculate its exact cost). We empirically showed that COINCIDE is sensitive to the layer choice. Table 5c compares the ARP of COINCIDE when varying the layer used to extract features. Clearly, the performance decreases significantly when using a different set of layers. Furthermore, COINCIDE requires substantially more memory: for each data point it stores a feature vector of size 20480, while XMAS only stores a vector of size 35.

Cluster diversity. As illustrated in Figure 6, the clusters identified by XMAS contain samples spanning multiple concepts detected by COINCIDE. This indicates that while samples within an XMAS cluster are conceptually diverse, they exhibit redundancy relative to each other in terms of training—an aspect that concept-based clustering methods like COINCIDE fails to capture. Figures 7-10 show the semantic diversity within a single XMAS cluster for different datasets. The values in the cross-model attention trajectory graphs also show that these clusters vary in the levels of cross-modal alignment.

Other target models. To assess the generalizability of XMAS, we train two additional target models—LLaVA-1.6-Mistral-7B and Phi-3.5-Vision-Instruct—using 10% subsets selected by the same proxy model (TinyLLaVA 2.0B), and compare XMAS against other methods. Table 6 shows that XMAS consistently achieves the best performance and is the only method that outperforms random selection across all target model architectures. In contrast, COINCIDE relies on hand-picked features, which fail to generalize to these new models.



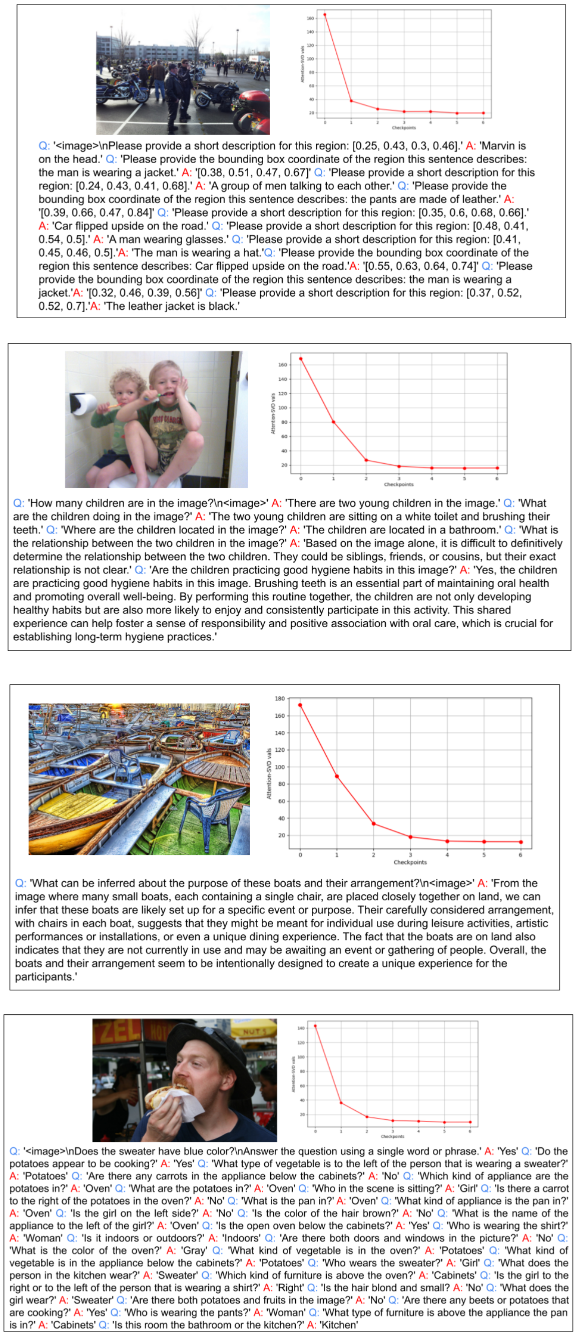
Q: "<image>\nWhat might have been the players' strategy during this game?" A: "In the image, there are three baseball players on the field: a pitcher, a baseman (likely the first baseman), and another fielder. Although it's not possible to ascertain their specific strategy without more context, it's likely that they are employing standard baseball tactics to maximize their team's defensive capabilities. The pitcher would be focused on delivering a variety of pitches to confuse the batter and attempt to get them out through strikes or by inducing weak contact. Meanwhile, the baseman would be ready to receive throws and make quick plays to potentially out runners trying to reach base. The other fielder would position himself to be alert and prepared to cover a substantial area of the field to catch or stop any incoming balls to prevent the opposing team from advancing on the bases. Overall, the players' strategy likely revolves around effective communication, anticipating the batter's tendencies, and working together as a cohesive unit to minimize the opposition's scoring opportunities."

Q: Please provide a short description for this region: [0.33, 0.47, 0.52, 0.71].' A: 'Chair left of coffee table.' Q: 'Please provide the bounding box coordinate of the region this sentence describes: chair near door.' A: '[0.33, 0.47, 0.52, 0.71]' Q: 'Please provide the bounding box coordinate of the region this sentence describes: chair next to lamp.' A: '[0.74, 0.53, 1.0, 0.86]' Q: 'Please provide the bounding box coordinate of the region this sentence describes: right chair.' A: '[0.74, 0.53, 1.0, 0.86]' Q: 'Please provide a short description for this region: [0.74, 0.53, 1.0, 0.86].' A: 'Red chair that is not all showing.' Q: 'Please provide a short description for this region: [0.33, 0.47, 0.52, 0.71].' A: 'A brown wicker chair with red cushions next to an open doorway.' Q: 'Please provide the bounding box coordinate of the region this sentence describes: armchair nearest the door.' A: '[0.33, 0.47, 0.52, 0.71]' Q: 'Please provide the bounding box coordinate of the region this sentence describes: chair furest from you.' A: '[0.33, 0.47, 0.52, 0.71]'

Q: What is the title of this book?\nAnswer the question using a single word or phrase.' A: 'Early Modern Social Theory: Selected Interpretive Readings' Q: 'What type of book is this?' A: 'Politics & Social Sciences' Q: 'Is this a sociopolitical book?' A: 'Yes' Q: 'Is this a life story book?' A: 'No'.

Q: Please provide a short description for this region: [0.38, 0.43, 0.51, 0.63].' A: 'Baby animal lying on ground.' Q: 'Please provide a short description for this region: [0.4, 0.45, 0.51, 0.58].' A: 'A giraffe laying down.' Q: 'Please provide the bounding box coordinate of the region this sentence describes: large giraffe standing up and looking right.' A: '[0.38, 0.21, 0.71, 0.49]' Q: 'Please provide the bounding box coordinate of the region this sentence describes: Small wooden log.' A: '[0.62, 0.13, 0.78, 0.19]' Q: 'Please provide a short description for this region: [0.02, 0.44, 0.29, 0.59].' A: 'The grass is green.' Q: 'Please provide a short description for this region: [0.37, 0.3, 0.43, 0.38].' A: "Large giraffe's tail that seems to be wagging." Q: 'Please provide the bounding box coordinate of the region this sentence describes: a wavy log stick out of mud.' A: '[0.66, 0.35, 1.0, 0.51]' Q: 'Please provide the bounding box coordinate of the region this sentence describes: dirt road.' A: '[0.01, 0.12, 0.75, 0.25]' Q: 'Please provide a short description for this region: [0.22, 0.34, 0.52, 0.58].' A: 'Small giraffes on the grass.' Q: 'Please provide the bounding box coordinate of the region this sentence describes: a dirt trail.' A: '[0.0, 0.13, 0.53, 0.24]'

Figure 7: Samples from the largest XMAS clusters in LLaVA 665k dataset.



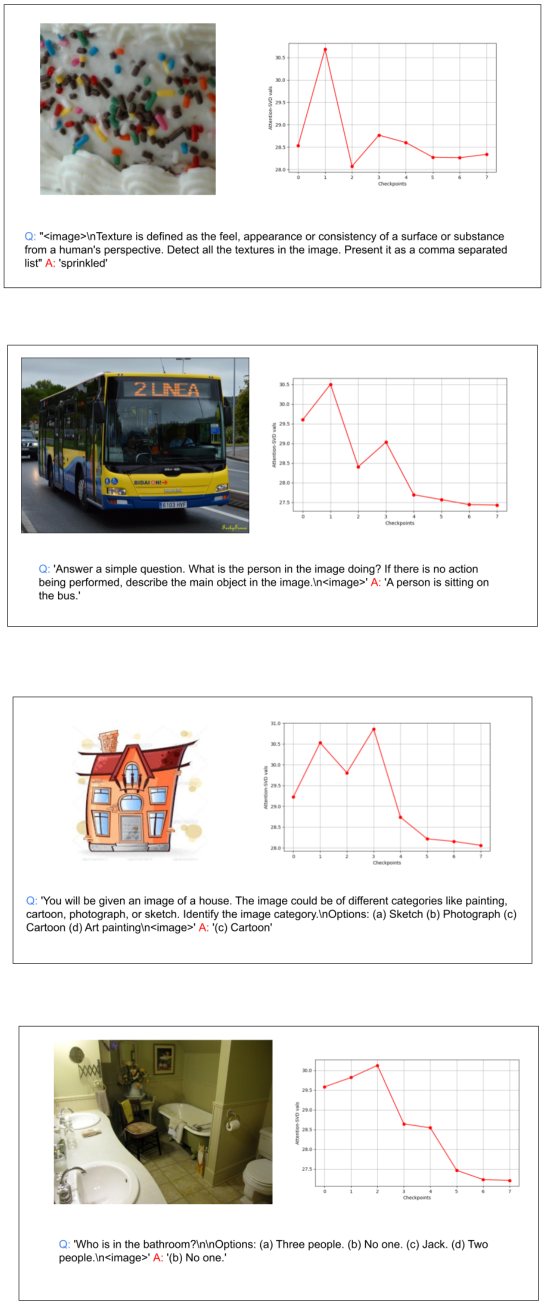
Q: '<image>\nPlease provide a short description for this region: [0.25, 0.43, 0.3, 0.46].' A: 'Marvin is on the head.' Q: 'Please provide the bounding box coordinate of the region this sentence describes: the man is wearing a jacket.' A: '[0.38, 0.51, 0.47, 0.67]' Q: 'Please provide a short description for this region: [0.24, 0.43, 0.41, 0.68].' A: 'A group of men talking to each other.' Q: 'Please provide the bounding box coordinate of the region this sentence describes: the pants are made of leather.' A: '[0.39, 0.66, 0.47, 0.84]' Q: 'Please provide a short description for this region: [0.35, 0.6, 0.68, 0.66].' A: 'Car flipped upside on the road.' Q: 'Please provide a short description for this region: [0.48, 0.41, 0.54, 0.5].' A: 'A man wearing glasses.' Q: 'Please provide a short description for this region: [0.41, 0.45, 0.46, 0.5].'A: 'The man is wearing a hat.'Q: 'Please provide the bounding box coordinate of the region this sentence describes: Car flipped upside on the road.'A: '[0.55, 0.63, 0.64, 0.74]' Q: 'Please provide the bounding box coordinate of the region this sentence describes: the man is wearing a jacket.'A: '[0.32, 0.46, 0.39, 0.56]' Q: 'Please provide a short description for this region: [0.37, 0.52, 0.52, 0.7].'A: 'The leather jacket is black.'

Q: 'How many children are in the image?\n<image>' A: 'There are two young children in the image.' Q: 'What are the children doing in the image?' A: 'The two young children are sitting on a white toilet and brushing their teeth.' Q: 'Where are the children located in the image?' A: 'The children are located in a bathroom.' Q: 'What is the relationship between the two children in the image?' A: 'Based on the image alone, it is difficult to definitively determine the relationship between the two children. They could be siblings, friends, or cousins, but their exact relationship is not clear.' Q: 'Are the children practicing good hygiene habits in this image?' A: 'Yes, the children are practicing good hygiene habits in this image. Brushing teeth is an essential part of maintaining oral health and promoting overall well-being. By performing this routine together, the children are not only developing healthy habits but are also more likely to enjoy and consistently participate in this activity. This shared experience can help foster a sense of responsibility and positive association with oral care, which is crucial for establishing long-term hygiene practices.'

Q: 'What can be inferred about the purpose of these boats and their arrangement?\n<image>' A: 'From the image where many small boats, each containing a single chair, are placed closely together on land, we can infer that these boats are likely set up for a specific event or purpose. Their carefully considered arrangement, with chairs in each boat, suggests that they might be meant for individual use during leisure activities, artistic performances or installations, or even a unique dining experience. The fact that the boats are on land also indicates that they are not currently in use and may be awaiting an event or gathering of people. Overall, the boats and their arrangement seem to be intentionally designed to create a unique experience for the participants.'

Q: '<image>\nDoes the sweater have blue color?\nAnswer the question using a single word or phrase.' A: 'Yes' Q: 'Do the potatoes appear to be cooking?' A: 'Yes' Q: 'What type of vegetable is to the left of the person that is wearing a sweater?' A: 'Potatoes' Q: 'Are there any carrots in the appliance below the cabinets?' A: 'No' Q: 'Which kind of appliance are the potatoes in?' A: 'Oven' Q: 'What are the potatoes in?' A: 'Oven' Q: 'Who in the scene is sitting?' A: 'Girl' Q: 'Is there a carrot to the right of the potatoes in the oven?' A: 'No' Q: 'What is the pan in?' A: 'Oven' Q: 'What kind of appliance is the pan in?' A: 'Oven' Q: 'Is the girl on the left side?' A: 'No' Q: 'Is the color of the hair brown?' A: 'No' Q: 'What is the name of the appliance to the left of the girl?' A: 'Oven' Q: 'Is the open oven below the cabinets?' A: 'Yes' Q: 'Who is wearing the shirt?' A: 'Woman' Q: 'Is it indoors or outdoors?' A: 'Indoors' Q: 'Are there both doors and windows in the picture?' A: 'No' Q: 'What is the color of the oven?' A: 'Gray' Q: 'What kind of vegetable is in the oven?' A: 'Potatoes' Q: 'What kind of vegetable is in the appliance below the cabinets?' A: 'Potatoes' Q: 'Who wears the sweater?' A: 'Girl' Q: 'What does the person in the kitchen wear?' A: 'Sweater' Q: 'Which kind of furniture is above the oven?' A: 'Cabinets' Q: 'Is the girl to the right or to the left of the person that is wearing a shirt?' A: 'Right' Q: 'Is the hair blond and small? A: 'No' Q: 'What does the girl wear?' A: 'Sweater' Q: 'Are there both potatoes and fruits in the image?' A: 'No' Q: 'Are there any beets or potatoes that are cooking?' A: 'Yes' Q: 'Who is wearing the pants?' A: 'Woman' Q: 'What type of furniture is above the appliance the pan is in?' A: 'Cabinets' Q: 'Is this room the bathroom or the kitchen?' A: 'Kitchen'

Figure 8: Samples from the smallest XMAS clusters in LLaVA 665k dataset.



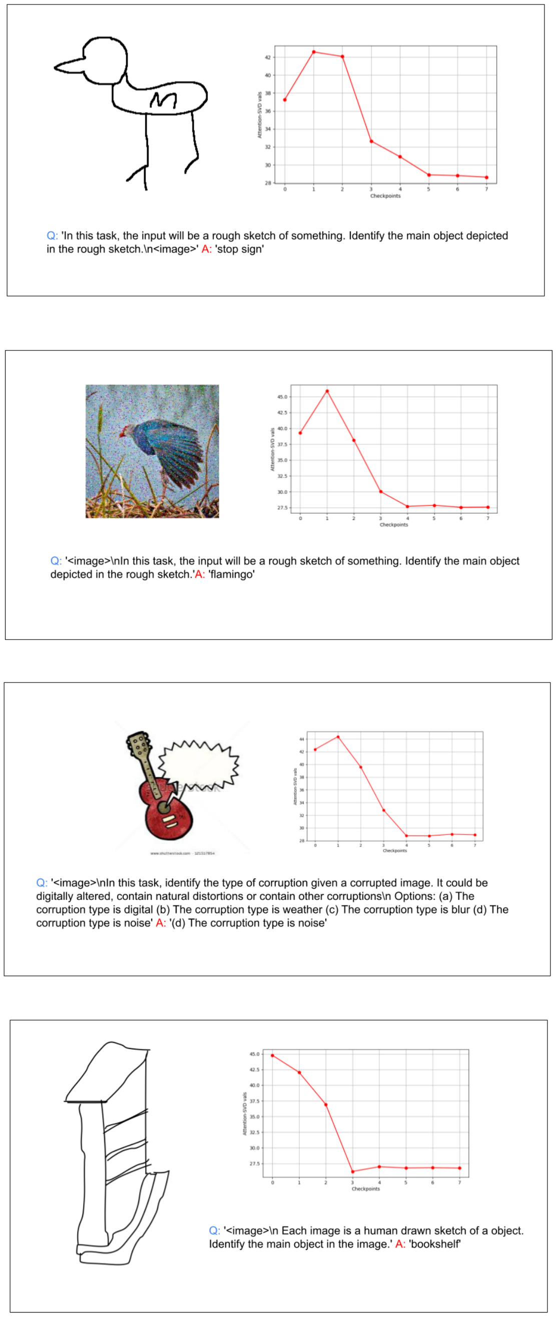
Q: "<image>\nTexture is defined as the feel, appearance or consistency of a surface or substance from a human's perspective. Detect all the textures in the image. Present it as a comma separated list" A: 'sprinkled'

Q: 'Answer a simple question. What is the person in the image doing? If there is no action being performed, describe the main object in the image.\n<image>' A: 'A person is sitting on the bus.'

Q: 'You will be given an image of a house. The image could be of different categories like painting, cartoon, photograph, or sketch. Identify the image category.\nOptions: (a) Sketch (b) Photograph (c) Cartoon (d) Art painting\n<image>' A: '(c) Cartoon'

Q: 'Who is in the bathroom?\n\nOptions: (a) Three people. (b) No one. (c) Jack. (d) Two people.\n<image>' A: '(b) No one.'

Figure 9: Samples from the largest XMAS clusters in Vision-Flan 191k dataset.



Q: 'In this task, the input will be a rough sketch of something. Identify the main object depicted in the rough sketch.\n<image>' A: 'stop sign'

Q: '<image>\nIn this task, the input will be a rough sketch of something. Identify the main object depicted in the rough sketch.'A: 'flamingo'

Q: '<image>\nIn this task, identify the type of corruption given a corrupted image. It could be digitally altered, contain natural distortions or contain other corruptions\n Options: (a) The corruption type is digital (b) The corruption type is weather (c) The corruption type is blur (d) The corruption type is noise' A: '(d) The corruption type is noise'

Q: '<image>\n Each image is a human drawn sketch of a object. Identify the main object in the image.' A: 'bookshelf'

Figure 10: Samples from the smallest XMAS clusters in Vision-Flan 191k dataset.