PRESERVE THE LANGUAGE, USE THE VISION: TRAINING-FREE DATA SELECTION FOR VISION-LANGUAGE INSTRUCTION TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual instruction tuning is essential for adapting vision-language models (VLMs) to vision-language tasks; however, this process may perturb the language representations acquired by their underlying large language models (LLMs) during language pretraining, thereby degrading their original language capabilities. Existing approaches primarily mitigate this issue through architectural modifications, additional text-only data, or training-time constraints, while largely overlooking the role of vision-language instruction data itself. In this work, we find that different vision-language samples induce substantially different degrees of perturbation to the original language representations of their underlying LLMs, suggesting that data selection can provide an effective way to mitigate language capability degradation. Based on this observation, we propose PLUV, a training-free vision-language data selection framework. PLUV first employs language forgetting risk estimation (LFRE) to select samples that induce less perturbation to the original language representations, and then uses visual dependency estimation (VDE) to measure the actual dependency of each sample on visual information. On text-only and vision-language benchmarks, PLUV achieves overall relative performance of 101.2% and 101.3%, respectively, demonstrating its ability to preserve language capabilities while maintaining vision-language performance and thereby achieving a better balance between the two.

1 INTRODUCTION

Vision-language models (VLMs) have demonstrated remarkable capabilities in processing and understanding vision-language data. Built upon large language models (LLMs), VLMs typically introduce a visual encoder to represent visual inputs and integrate the resulting visual features with textual inputs, enabling joint modeling and complex interaction across modalities (Liu et al., 2023; Li et al., 2023a). However, while vision-language adaptation substantially enhances the vision-language capabilities of VLMs, it can also degrade their original text-only capabilities to some extent (Li et al., 2024; Yu & Ananiadou, 2025). As shown in Figure 1, VLMs achieve strong performance on vision-language benchmarks, yet often exhibit performance degradation on text-only tasks compared with their underlying LLMs.

Existing approaches mainly address this issue by scaling up model size, incorporating additional high-quality text-only data, or freezing part of the LLMs backbone (Dai et al., 2024; Lin et al., 2024; Wang et al., 2025). However, these strategies often struggle to balance vision-language performance and training efficiency. Scaling model size or introducing additional data substantially increases computational and data costs (Dai et al., 2024; Lin et al., 2024), while freezing part of the model parameters may limit the model's ability to adapt to vision-language tasks (Wang et al., 2025). This motivates a more appealing alternative: rather than modifying the model architecture or relying on additional training data, we seek to identify more suitable samples directly from the existing vision-language instruction-tuning data, thereby improving vision-language capabilities while preserving the language knowledge and reasoning abilities inherited from the underlying LLMs.

As shown in Appendix A, different vision-language samples exhibit substantially different effects on textual representations after visual information is introduced. For some samples, the textual at-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

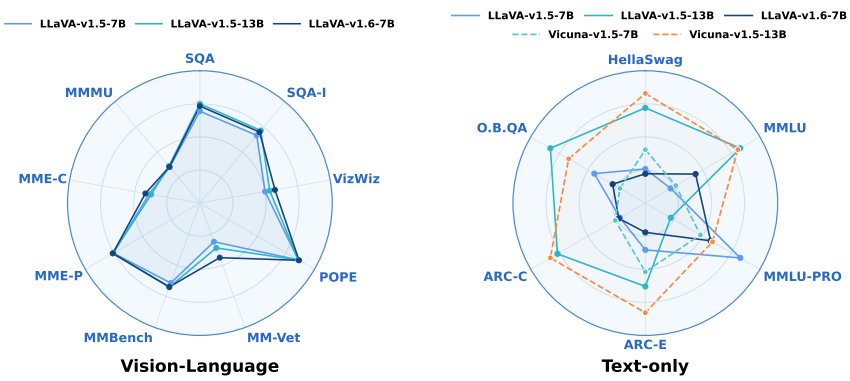


Figure 1: Comparison of vision-language and text-only capabilities. **(Left)** Performance of VLMs on vision-language benchmarks. **(Right)** Performance of VLMs and their underlying LLMs on text-only benchmarks.

tention representations under vision-language and text-only inputs remain highly consistent, with similarities exceeding 0.7, suggesting that the introduction of visual information does not substantially alter the model's original text modeling patterns. In other words, these samples induce only limited perturbation to the language representations. In contrast, other samples exhibit considerably lower similarity, indicating that visual information more strongly alters the original textual attention structure and leads to larger representational shifts. These substantial sample-level differences suggest that vision-language data do not affect the original text capabilities of an VLMs to the same extent. Based on this observation, we argue that the consistency of internal textual representations before and after introducing visual information can serve as an effective signal for measuring the extent to which a vision-language sample perturbs the underlying LLMs.

To this end, we propose **PLUV**, a vision-language data selection framework that evaluates training samples from two complementary perspectives: language representation preservation and visual information utilization. Specifically, PLUV consists of two stages: Language Forgetting Risk Estimation (LFRE) and Visual Dependency Estimation (VDE). First, LFRE constructs two input views for each sample, namely a vision-language view and a text-only view, and compares the dominant singular components of the text-to-text attention structures under the two conditions to quantify the extent to which a sample perturbs the original language representations of the VLMs. Second, to avoid favoring samples whose language representations remain stable but that do not actually rely on visual information, VDE randomly ablates the image-to-text information flow within attention heads and estimates the visual contribution of each attention head according to the resulting change in the correct-answer score. This allows us to measure the extent to which each sample genuinely depends on visual information. By jointly considering these two dimensions, PLUV identifies high-quality vision-language samples that preserve language representations while effectively exploiting visual information, thereby mitigating language capability degradation during visual instruction tuning.

- We reveal that different vision-language samples induce substantially different levels of perturbation to the original language representations of VLMs. We further propose using the consistency of textual attention representations between vision-language and text-only inputs to characterize such sample-level language perturbation, providing a new data-centric perspective for understanding and mitigating language capability degradation during vision-language instruction tuning.
- We propose PLUV, a training-free vision-language data selection framework. PLUV first employs LFRE to quantify changes in language representations caused by the introduction of visual information, and then uses VDE to evaluate the genuine visual dependency of each sample through attention-head-level ablation of visual information flow, enabling the selection of high-quality vision-language training data.
- Experiments demonstrate that PLUV achieves a better balance between vision-language capability and language capability preservation using substantially less training data, validating the effectiveness of our proposed vision-language data selection framework.

2 Related Work

Visual Instruction Tuning. Visual instruction tuning has become a central paradigm for extending pretrained LLMs to vision-language scenarios (Liu et al., 2023; Li et al., 2023a; Bai et al., 2025). By constructing visual instruction data from images, textual instructions, and corresponding responses, and subsequently fine-tuning models on such data, VLMs acquire capabilities in image-text understanding, visual question answering, and visual reasoning. Recent works such as LLaVA (Liu et al., 2023), InstructBLIP (Li et al., 2023a), and Qwen-VL (Bai et al., 2025) have further advanced the extension of LLMs from natural language processing to vision-language applications. Despite its effectiveness in improving vision-language capabilities, visual instruction tuning inevitably alters the representation space and parameter distribution of the underlying LLMs, which may in turn degrade their pre-existing text-only capabilities (Lu et al., 2025; Chen et al., 2026).

Visual Instruction Data Selection. As visual instruction datasets continue to grow in scale, data redundancy has become increasingly prominent, resulting in substantial training costs. To address this issue, a growing body of work has explored visual instruction data selection, aiming to identify a small subset of high-value samples from large-scale datasets while achieving performance comparable to, or even better than, full-data fine-tuning at a lower training cost (Liu et al., 2025; Yu et al., 2025a; Bi et al., 2025; Safaei et al., 2025; Yang et al., 2025). Existing approaches can be broadly categorized into proxy-model-based selection methods and training-signal-based selection methods. Although these methods have made notable progress in improving training efficiency, their selection objectives primarily focus on vision-language performance, sample representativeness, or training efficiency. In other words, they mainly ask which samples are most beneficial for improving vision-language capabilities, while paying limited attention to whether these samples interfere with the pre-existing language capabilities of the underlying LLMs during vision-language adaptation.

Mitigating Language Capability Degradation in VLMs. Existing approaches for mitigating language capability degradation in VLMs can be broadly grouped into three categories (Team, 2024; Tu et al., 2024; Yu et al., 2024; 2025b; Chen et al., 2025). The first category freezes LLMs parameters, fixing part or all of the LLMs backbone to reduce the perturbation of pretrained language knowledge during vision-language training. However, this strategy may also restrict the model's ability to fully adapt to vision-language tasks (Dai et al., 2024; Team, 2024). The second category incorporates text-only data during vision-language training to help preserve language capabilities, but this introduces additional data and training costs (Lin et al., 2024; Tu et al., 2024). The third category employs parameter constraints or parameter-efficient fine-tuning, such as LoRA, adapters, and regularization, to limit the extent of parameter updates and thereby alleviate language forgetting. Nevertheless, these methods still rely on additional constraints imposed during the training process (Yu et al., 2024; 2025b; Chen et al., 2025). In contrast, our work approaches the problem from the perspective of the visual instruction data itself. We investigate how different vision-language samples perturb the language representations of the underlying LLMs and mitigate language capability degradation through data selection.

3 Methodology

Given a vision-language instruction data pool ($\mathcal{D}_{v\tau} = \{d_i = (x_i, v_i, y_i)\}_{i=1}^{N}$), where x_i denotes the textual instruction, v_i denotes the corresponding image, and y_i denotes the target answer, our goal is to select high-quality samples from the original data pool that simultaneously preserve language capabilities and provide effective visual learning signals. As illustrated in Figure 2, we propose a two-stage data selection framework consisting of Language Forgetting Risk Estimation (LFRE) and Visual Dependency Estimation (VDE). First, LFRE estimates the perturbation to language representations caused by the introduction of visual information by comparing the internal text-to-text attention structures of the same sample under vision-language and text-only input conditions. Based on this comparison, LFRE assesses the potential language forgetting risk of each sample and selects samples that have a smaller impact on the model's original language capabilities, thereby forming a candidate set $\mathcal{D}_{LFRE}$ with relatively low language forgetting risk. Subsequently, VDE further estimates the actual dependency of these candidate samples on visual information through attention-head-level ablation of visual information flow. Finally, we prioritize samples that induce relatively small perturbations to language representations while exhibiting strong visual dependency, yielding the selected subset $\mathcal{D}_{PLUV}$ for visual instruction tuning.

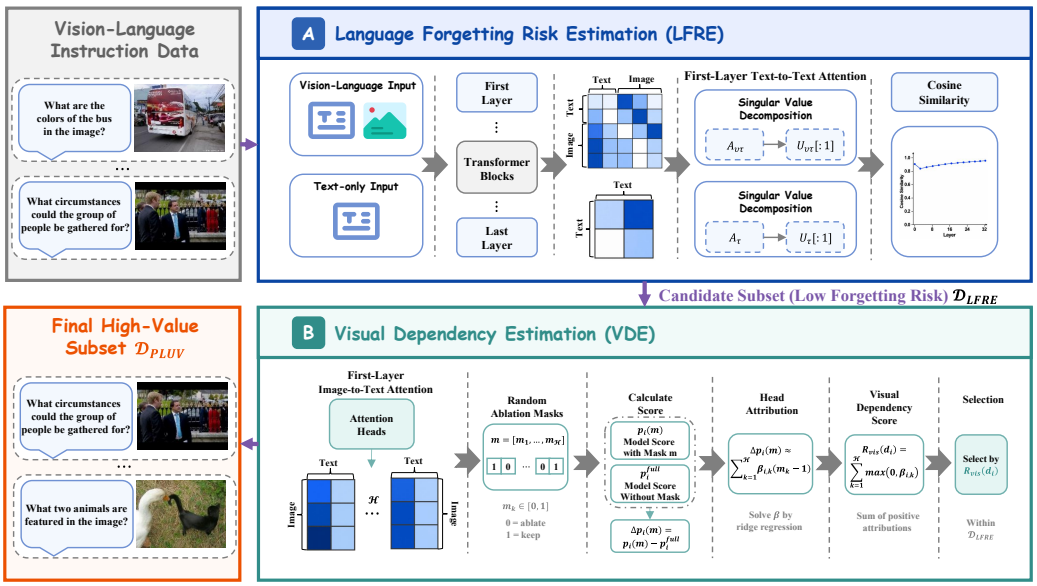


Figure 2: Overview of the proposed PLUV framework. PLUV sequentially applies **Language Forgetting Risk Estimation (LFRE)** and **Visual Dependency Estimation (VDE)** to select high-value vision-language data that preserve language capabilities while utilizing visual information.

3.1 Language Forgetting Risk Estimation (LFRE)

To estimate the potential impact of vision-language samples on the original language capabilities of LLMs, as shown in Figure 2(a), LFRE first construct two corresponding input views for each sample d_i: $v_{\upsilon\tau}(d_i) = (x_i, v_i)$ and $v_\tau(d_i) = x_i$, where $v_{\upsilon\tau}(d_i)$ denotes the vision-language view and $v_\tau(d_i)$ denotes the text-only view. The two views share exactly the same textual content and therefore differ only in the presence of visual information. We denote the pretrained language model before visual instruction tuning as M. Throughout the data evaluation process, the parameters of M remain frozen and are not updated by gradients. The model is used solely as a fixed language model for analyzing changes in its internal representations under different input conditions.

We find that the attention structure in the first layer exhibits strong consistency with those in deeper layers, suggesting that the first layer alone is sufficient to effectively characterize the structural differences between the two views. A detailed analysis is provided in Appendix A.1. Based on this observation, for the first transformer layer of M, we extract the attention matrices among textual tokens from the vision-language and text-only views. Since transformers layers employ multi-head attention, we average the text-to-text attention submatrices across all attention heads, denoted by $A_{i,\upsilon\tau}$ and $A_{i,\tau}$, respectively. Importantly, for the vision-language view, only the text-to-text attention submatrix is retained from the full attention matrix. Therefore, both $A_{i,\upsilon\tau}$ and $A_{i,\tau}$ characterize the internal dependencies among textual tokens, with the key difference that $A_{i,\upsilon\tau}$ is obtained in the presence of visual information.

Directly comparing the complete attention matrices may be sensitive to local noise and low-energy components. To capture their dominant structures, we perform singular value decomposition (SVD) on the attention matrices from the two views:

$$A_{i,j} = U_{i,j}\Sigma_{i,j}(V_{i,j})^\top, \qquad j \in \{\upsilon\tau, \tau\}. \tag{1}$$

We further find that the leading singular component is sufficient to effectively characterize the dominant structure of the attention matrix, indicating that using only this component can preserve its primary structural information. A detailed analysis is provided in Appendix A.2. Therefore, we use the singular vector corresponding to the largest singular value to represent the dominant structure of the attention matrix. Specifically, we define $u_{i,\upsilon\tau} = U_{i,\upsilon\tau}[:, 1]$ and $u_{i,\tau} = U_{i,\tau}[:, 1]$. We then compute the cosine similarity between the dominant attention structures of the two views as:

$$S_i = \cos\left(u_{i,\upsilon\tau}, u_{i,\tau}\right). \tag{2}$$

A higher S_i indicates that the dominant internal attention structure among textual tokens remains more consistent with that under the original text-only condition after visual information is introduced. In contrast, a lower S_i indicates that visual information causes a stronger alteration to the original textual dependency structure. Therefore, samples with higher S_i, corresponding to lower potential language forgetting risk, are selected to construct the candidate set $\mathcal{D}_{LFRE}$. As shown in Figure 2(a), LFRE retains samples with relatively stable text representations to reduce the potential interference of visual instruction tuning with the original language capabilities of LLMs.

3.2 Visual Dependency Estimation (VDE)

Selecting samples solely according to language representation stability introduces a potential issue: the resulting subset may favor samples that are language-friendly but do not actually require visual information. For example, when a question contains strong textual priors, VLMs may still predict the correct answer even when visual information is completely ignored. Therefore, we further estimate the actual contribution of visual information to answer prediction for each candidate sample.

Rather than directly relying on attention weights, we focus on identifying which attention heads in the first transformer layer actually transmit visual information that is useful for answering the current question from image tokens to textual tokens (Kang et al., 2025a). Suppose the first transformer layer contains $\mathcal{H}$ attention heads to be analyzed. We denote the k-th attention head by $\mathcal{H}_k$, where k corresponds to the head index within this layer. For a sample d_i, we collectively refer to all attention pathways from image tokens to textual tokens within $\mathcal{H}_k$ as the visual information flow of that head. Individually ablating every attention head would provide a direct estimate of its contribution, but this becomes computationally expensive when the model contains a large number of attention heads (Kang et al., 2025b). Therefore, as shown in Figure 2(b), VDE adopts a random ablation strategy, which enables the contributions of multiple attention heads to be estimated using a limited number of forward passes. For each sample d_i, we construct a binary mask $m = [m_1, m_2, \ldots, m_{\mathcal{H}}]$, where:

$$m_k = \begin{cases} 1, & \text{retain the image-to-text information flow of } \mathcal{H}_k, \\ 0, & \text{ablate the image-to-text information flow of } \mathcal{H}_k. \end{cases} \tag{3}$$

By randomly sampling different masks m, multiple attention heads can be perturbed simultaneously within a single forward pass, allowing us to efficiently characterize the relationship between different combinations of attention-head ablations and changes in model predictions. Let the target answer of sample d_i consist of R_i tokens, $Y_i = \{y_{i,1}, y_{i,2}, \ldots, y_{i,R_i}\}$. Given a random ablation mask $\mathbf{m}$, we compute the average token-level log-likelihood assigned by the model to the target answer:

$$p_i(m) = \frac{1}{R_i} \sum_{r=1}^{R_i} \log P\left(y_{i,r} \mid v_i, x_i, y_{i,<r}; m\right). \tag{4}$$

When all image-to-text information flows are retained, we denote the answer score of the full model as $p_i^{\text{full}} = p_i(m_{\text{full}})$. For an arbitrary random mask m, we define the score change relative to the full model as:

$$\Delta p_i(m) = p_i(m) - p_i^{\text{full}}. \tag{5}$$

If ablating a particular set of visual information flows results in a substantial decrease in $\Delta p_i(m)$, it indicates that the removed attention heads transmit visual information that is beneficial for predicting the correct answer. For each sample d_i, we randomly generate Z joint ablation masks $\left\{m^{(1)}, m^{(2)}, \ldots, m^{(Z)}\right\}$ and obtain the corresponding score changes $\left\{\Delta p_i^{(1)}, \Delta p_i^{(2)}, \ldots, \Delta p_i^{(Z)}\right\}$ through forward passes. To recover the marginal contribution of individual attention heads from the joint ablation results, we employ an additive linear model to approximate the effect of visual information flow from different attention heads on the answer score:

$$\Delta p_i(m) \approx \sum_{k=1}^{\mathcal{H}} \beta_{i,k}(m_k - 1). \tag{6}$$

Here, $m_k - 1 = 0$ indicates that the corresponding attention head is retained, whereas $m_k - 1 = -1$ indicates that its image-to-text information flow is ablated. We therefore estimate the individual contribution of each attention head from the joint ablation results using ridge regression:

$$\beta_i = \arg\min_{\beta} \sum_{z=1}^{Z} \left[\Delta p_i^{(z)} - \sum_{k=1}^{\mathcal{H}} \beta_k \left(m_k^{(z)} - 1\right)\right]^2 + \lambda \|\beta\|_2^2. \tag{7}$$

This yields the sample-level attention-head attribution vector $\beta_i = [\beta_{i,1}, \beta_{i,2}, \ldots, \beta_{i,\mathcal{H}}]$. Here, $\beta_{i,k}$ measures the contribution of the visual information transmitted by the k-th attention head to the answer score of sample d_i. A larger $\beta_{i,k}$ indicates that ablating the corresponding image-to-text information flow causes a larger decrease in the correct-answer score, suggesting that the attention head plays a more important role in utilizing visual information for the current sample. Finally, we define the visual dependency score of sample d_i as:

$$R_{\text{vis}}(d_i) = \sum_{k=1}^{\mathcal{H}} \max\left(0, \beta_{i,k}\right). \tag{8}$$

A larger $R_{\text{vis}}(d_i)$ indicates that disrupting the image-to-text information flow leads to a more substantial degradation in the model's ability to predict the correct answer, suggesting that the sample exhibits stronger dependency on visual information. As shown in Figure 2(b), samples in the candidate set $\mathcal{D}_{LFRE}$ are ranked according to R_{vis}, and those with higher visual dependency scores are retained. After the two-stage selection based on language forgetting risk and visual dependency, we obtain the final high-value subset $\mathcal{D}_{PLUV}$, which is subsequently used for visual instruction tuning.

4 Experiments

4.1 Experiment Setup

Dataset & Model. We evaluate **PLUV** on two visual instruction-tuning datasets, **LLaVA-665K** (Liu et al., 2024a) and **VisionFlan-186K** (Xu et al., 2024), with **LLaVA-v1.5-7B** (Liu et al., 2024a) serving as our primary base model. All models are trained for **one epoch** following the official fine-tuning configurations. To ensure a fair comparison across different methods, we use the same training configuration and experimental environment for all experiments.

Benchmarks. To comprehensively evaluate PLUV in terms of both text-only and vision-language capabilities, we conduct evaluations on a diverse set of text-only and vision-language benchmarks. For text-only evaluation, we use HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021), MMLU-Pro (Wang et al., 2024), ARC (Clark et al., 2018), and OpenBookQA (Mihaylov et al., 2018) to assess the model's language capabilities. For vision-language evaluation, we consider ScienceQA (Lu et al., 2022), VizWiz (Gurari et al., 2018), POPE (Li et al., 2023b), MMBench (Liu et al., 2024b), MME (Fu et al., 2025), and MMMU (Yue et al., 2024). All text-only benchmarks are evaluated using the **lm-eval-harness** framework (Gao et al., 2024), while the vision-language benchmarks are evaluated using **lmms-eval** (Zhang et al., 2025). Detailed descriptions of all benchmarks are provided in Appendix B.

Baselines. For the text-only benchmarks, we compare PLUV against random selection, COINCIDE (Lee et al., 2024), DataTailor (Yu et al., 2025a) and PRISM (Bi et al., 2025) as representative baselines to assess how different data selection strategies preserve the original language capabilities of the underlying LLMs during visual instruction tuning. For the vision-language benchmarks, we conduct a comprehensive comparison with a broad range of representative data selection baselines, including random selection, Instruction Length, Perplexity (Marion et al., 2023), GraNd (Paul et al., 2021), EL2N (Paul et al., 2021), InstructionGPT-4 (Wei et al., 2023), SELF-FILTER (Chen et al., 2024), TIVE (Liu et al., 2025), COINCIDE (Lee et al., 2024), DataTailor (Yu et al., 2025a), ICONS (Wu et al., 2025), and PRISM (Bi et al., 2025).

4.2 Experimental Results

Under the same vision-language instruction tuning and evaluation settings, we compare PLUV with representative data selection baselines. For PLUV, we set the similarity threshold in LFRE to 0.7 and the retained ratio in VDE to 0.4, resulting in a final selection of 72128 samples. As shown in Table 1, existing data selection baselines may lead to varying degrees of language capability degradation after vision-language instruction tuning, whereas PLUV achieves an overall relative performance (Rel.) of 101.2% on text-only benchmarks, demonstrating that the selected data can better preserve the model's original language capability. Moreover, as shown in Table 2, PLUV maintains vision-language understanding capability, achieving an overall relative performance of 101.3% on vision-language benchmarks. These results indicate that PLUV does not preserve language capability at

the expense of vision-language learning. Instead, it selects samples that exhibit both low language forgetting risk and visual dependency, thereby achieving an effective balance between language preservation and vision-language instruction tuning.

Table 1: Performance comparison of PLUV with Full-Finetune and existing data selection approaches on text-only benchmarks. Full-Finetune is normalized to 100%. The Vicuna-v1.5-7B language model is also included as the base language model for reference.

Methods	HellaSwag ↑	MMLU ↑	MMLU-PRO ↑	ARC-E ↑	ARC-C ↑	O.B.QA ↑	Rel. (%)
Base Language Model							
Vicuna-v1.5-7B	76.9	50.9	11.6	80.3	50.1	45.5	
Full-Finetune	75.7	50.5	15.9	78.8	49.7	46.8	100%
Random	74.7	49.9	9.1	79.0	47.8	46.2	91.6%
COINCIDE	75.5	49.9	11.6	79.8	50.1	44.2	94.7%
DataTailor	75.7	50.1	14.5	78.5	49.7	46.3	98.2%
PRISM	75.4	49.2	9.3	78.2	50.3	44.6	91.9%
PLUV	**75.8**	**50.0**	**17.3**	**79.6**	**50.3**	**45.4**	**101.2%**

Table 2: Performance comparison of PLUV with Full-Finetune and existing data selection approaches on vision-language benchmarks. Full-Finetune normalized to 100%. "–" indicates that the corresponding result is not reported.

Methods	SQA ↑	SQA-I ↑	VizWiz ↑	POPE ↑	MMBench ↑	MME-P ↑	MME-C ↑	MMMU ↑	Rel. (%)
Full-Finetune	69.4	66.8	50.0	85.9	64.3	1510.7	311.9	35.4	100%
Random	65.5	64.5	48.1	84.4	55.5	1492.0	233.5	30.5	91.4%
Length	66.8	66.7	47.0	85.0	57.0	1422.1	306.0	33.1	95.4%
EL2N	70.2	70.6	44.4	85.6	61.6	1356.5	294.7	–	96.5%
Perplexity	70.5	67.9	–	83.3	62.3	1393.3	260.7	–	95.5%
GraNd	71.4	68.4	37.8	82.5	62.9	1400.5	287.1	–	94.2%
TIVE	72.2	70.6	–	85.6	63.2	1433.0	322.1	–	101.0%
InstructionGPT-4	–	–	–	–	31.4	463.3	–	–	39.8%
Self-Filter	–	61.4	53.2	83.8	61.4	1306.2	–	–	95.6%
COINCIDE	–	69.2	46.8	86.1	63.1	1495.6	–	–	98.9%
ICONS	–	70.8	–	87.5	63.1	1485.7	–	–	101.1%
DataTailor	71.0	–	49.5	85.3	–	1476.1	319.2	–	100.1%
PRISM	71.3	69.1	50.1	87.3	65.2	1470.0	330.0	34.7	101.3%
PLUV	**70.3**	**69.3**	**52.0**	**87.3**	**65.4**	**1432.8**	**328.2**	**34.8**	**101.3%**

4.3 Generalization across Different Data Pools

To evaluate whether PLUV can generalize beyond the original instruction data pool, we further conduct experiments on another data pool, VisionFlan-186K. Under the same evaluation setting, we compare PLUV with full-data training, random selection, and PRISM. As shown in Figure 3, PLUV maintains performance on both text-only and vision-language benchmarks using only 28K samples. Detailed results on each evaluation benchmark are provided in Appendix C. On text-only benchmarks, PLUV achieves an overall relative performance of 102.1%, indicating that the selected vision-language data effectively preserve the language capabilities of VLMs. On vision-language benchmarks, PLUV similarly achieves an overall relative performance of 101.2%, further demonstrating that the selected data maintain the vision-language capabilities of VLMs. These results indicate that PLUV generalizes effectively across different data pools.

4.4 Computational Cost Analysis

To evaluate the practical computational cost of different data selection baselines, we report the runtime of the entire pipeline in Figure 4. All experiments are conducted on two NVIDIA A6000 48GB GPUs, and identical training configurations are adopted during the vision-language instruction tuning stage to ensure a fair comparison. The total runtime is divided into two stages: data selection and vision-language instruction tuning. Full-Finetune does not require an additional data selection stage, but training on the full dataset takes 71 hours. In comparison, COINCIDE, DataTailor, PRISM, and

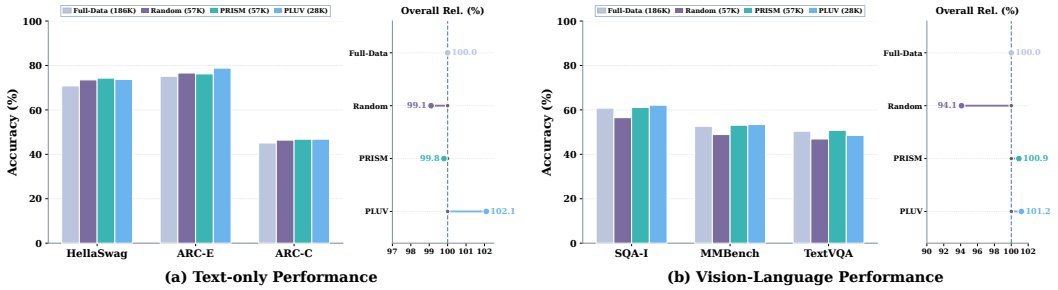


Figure 3: Generalization performance on the VisionFlan-186K data pool. Performance comparisons among Full-Data, Random, PRISM, and PLUV on selected text-only and vision-language benchmarks, together with overall relative performance reported for both text-only and vision-language benchmarks. Detailed results are provided in Tables 7 and 8. (a) Performance on text-only benchmarks. (b) Performance on vision-language benchmarks.

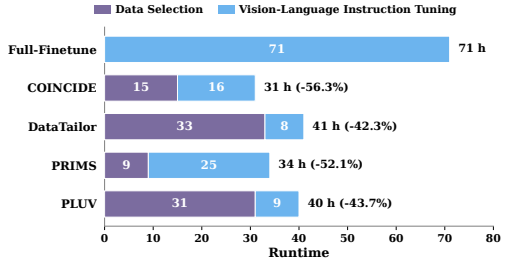


Figure 4: Runtime comparison of different data selection baselines. Runtime is reported in hours and decomposed into data selection and vision-language instruction tuning. The percentages in parentheses indicate the reduction in total runtime relative to Full-Finetune.

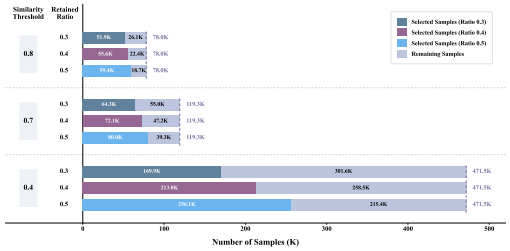


Figure 5: Number of samples under different similarity thresholds and retained ratios. The bars show the total number of samples in the candidate set $\mathcal{D}_{LFRE}$ and the number of samples retained in the final subset $\mathcal{D}_{PLUV}$ under different settings.

PLUV reduce the overall runtime by 56.3%, 42.3%, 52.1%, and 43.7%, respectively, indicating that data selection can effectively reduce the overall computational cost to varying degrees during the instruction tuning stage. A more detailed analysis of the computational cost of PLUV during the data selection stage is provided in Appendix D.

4.5 Effect of Selection Threshold and Retained Ratio

We further investigate the effects of the similarity threshold in LFRE and the retained ratio in VDE on the performance of PLUV. As shown in Figure 5, increasing the similarity threshold results in a smaller candidate subset $\mathcal{D}_{LFRE}$, while the retained ratio directly determines the size of the final subset $\mathcal{D}_{PLUV}$. Different combinations of thresholds and retained ratios provide varying selection granularities, allowing PLUV to adapt to different data scale requirements while maintaining effective sample selection. As shown in Figures 5 and 6, although different configurations lead to some variation in both the number of retained samples and the overall relative performance, PLUV maintains relatively stable performance across a wide range of parameter settings, and achieves performance comparable to or even exceeding Full-Finetune under certain configurations. Detailed results on both text-only and vision-language benchmarks are provided in Appendix E.

4.6 Ablation Study

Effectiveness of LFRE. To evaluate the effectiveness of LFRE, we fine-tune the model using only the subset $\mathcal{D}_{LFRE}$ selected by LFRE from the full data pool. As shown in Tables 3 and 4, when the similarity threshold is set to 0.7, the resulting subset $\mathcal{D}_{LFRE}$ contains only 119287 samples. Compared with the subset removed by LFRE, $\mathcal{D}_{vT} \backslash \mathcal{D}_{LFRE}$, $\mathcal{D}_{LFRE}$ improves the overall relative

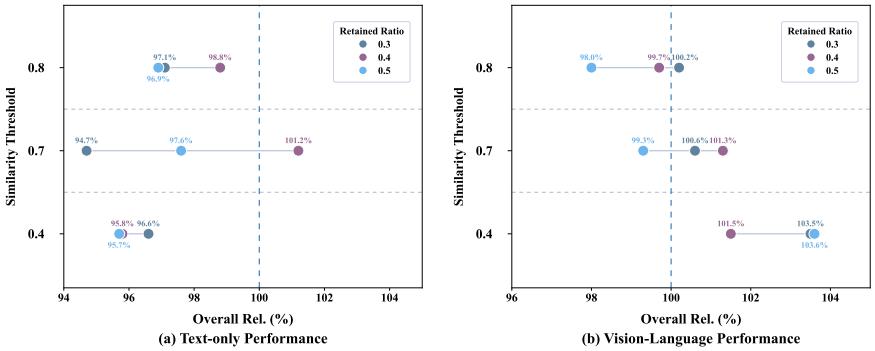


Figure 6: Overall relative performance under different similarity thresholds and retained ratios. Points with different colors denote different retained ratios, and the vertical dashed line indicates 100% relative performance. Detailed results are provided in Tables 10 and 11. (a) Performance on text-only benchmarks. (b) Performance on vision-language benchmarks.

Table 3: Text-only performance of different subsets selected at each stage of PLUV.

Selection Stage	Subset	Samples	HellaSwag ↑	MMLU ↑	MMLU-PRO ↑	ARC-E ↑	ARC-C ↑	O.B.QA ↑	Rel. (%)
Full-Finetune	$\mathcal{D}_{v\tau}$	665298	75.7	50.5	15.9	78.8	49.7	46.8	100%
LFRE Selected	$\mathcal{D}_{LFRE}$	119287	75.8	50.0	14.3	79.1	49.3	45.6	97.7%
LFRE Unselected	$\mathcal{D}_{v\tau} \setminus \mathcal{D}_{LFRE}$	546011	73.5	49.1	11.3	77.0	49.2	47.1	93.8%
PLUV Selected	$\boldsymbol{\mathcal{D}_{PLUV}}$	**72128**	**75.8**	**50.0**	**17.3**	**79.6**	**50.3**	**45.4**	**101.2%**

Table 4: Vision-language performance of different subsets selected at each stage of PLUV.

Selection Stage	Subset	Samples	SQA ↑	SQA-I ↑	VizWiz ↑	POPE ↑	MMBench ↑	MME-P ↑	MME-C ↑	MMMU ↑	Rel. (%)
Full-Finetune	$\mathcal{D}_{v\tau}$	665298	69.4	66.8	50.0	85.9	64.3	1510.7	311.9	35.4	100%
LFRE Selected	$\mathcal{D}_{LFRE}$	119287	70.0	68.1	51.1	87.2	65.3	1442.4	322.9	35.0	100.7%
LFRE Unselected	$\mathcal{D}_{v\tau} \setminus \mathcal{D}_{LFRE}$	546011	67.7	65.9	51.5	87.3	64.1	1440.6	307.9	34.7	99.1%
PLUV Selected	$\boldsymbol{\mathcal{D}_{PLUV}}$	**72128**	**70.3**	**69.3**	**52.0**	**87.3**	**65.4**	**1432.8**	**328.2**	**34.8**	**101.3%**

performance on the text-only and vision-language benchmarks by 3.9% and 1.6%, respectively. These results indicate that the samples filtered out by LFRE tend to introduce larger perturbations to the model's language representations, thereby degrading the overall relative performance. This demonstrates the effectiveness of LFRE in identifying samples with lower language forgetting risk.

Effectiveness of VDE. To evaluate the effectiveness of VDE, we further apply VDE to $\mathcal{D}_{LFRE}$ with a retained ratio of 0.4, yielding the final subset $\mathcal{D}_{PLUV}$. As shown in Tables 3 and 4, compared with $\mathcal{D}_{LFRE}$, which contains 119287 samples, $\mathcal{D}_{PLUV}$ contains only 72128 samples, while achieving improvements of 3.5% and 0.6% in overall relative performance on the text-only and vision-language benchmarks, respectively. These results show that VDE can further identify samples with stronger dependence on visual information from the candidate set with low language forgetting risk, thereby improving the effectiveness of the selected data.

5 Conclusion

In this paper, we find that different vision-language instruction samples induce varying degrees of perturbation to the language representations of VLMs. Based on this observation, we propose PLUV, a training-free vision-language data selection framework that jointly considers language representation preservation and visual information utilization through LFRE and VDE. Specifically, LFRE identifies samples that introduce perturbations to language representations, while VDE further selects samples with stronger dependencies on visual information. Compared with existing data selection baselines, PLUV more effectively mitigates the degradation of VLMs' language capabilities during vision-language instruction tuning while preserving their vision-language performance.

Reproducibility Statement

We have made efforts to ensure the reproducibility of our experiments. The complete methodology of PLUV, including LFRE and VDE, is described in Section 3. The experimental setup, including the data pools, model configurations, baseline, training settings, and evaluation protocols, is provided in Section 4. Additional implementation details, hyperparameter settings, analyses of design choices, and computational costs are provided in the appendix. We also report the dataset sizes and selection configurations used in our experiments to facilitate reproduction of the selected subsets and experimental results. The implementation code and experiment configurations are provided as anonymized supplementary materials.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025.

Jinhe Bi, Yifan Wang, Danqi Yan, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. PRISM: self-pruning intrinsic selection method for training-free multimodal data selection. *CoRR*, abs/2502.12119, 2025.

Jinpeng Chen, Runmin Cong, Yuzhi Zhao, Hongzheng Yang, Guangneng Hu, Horace H. S. Ip, and Sam Kwong. SEFE: superficial and essential forgetting eliminator for multimodal continual instruction tuning. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

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. In *ACL (Findings)*, volume ACL 2024 of *Findings of ACL*, pp. 4156–4172. Association for Computational Linguistics, 2024.

Zixuan Chen, Juncheng Tao, and Ziqian Zeng. Preserving language capabilities in vision-language models via representation regulation. In *ACL (Findings)*, pp. 24189–24205. Association for Computational Linguistics, 2026.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457, 2018.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuoling Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NVLM: open frontier-class multimodal llms. *CoRR*, abs/2409.11402, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *NeurIPS*, 2025.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 07 2024. URL https://zenodo.org/records/12608602.

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 *CVPR*, pp. 3608–3617. Computer Vision Foundation / IEEE Computer Society, 2018.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*. OpenReview.net, 2021.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *ICLR*. OpenReview.net, 2025a.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs A few attention heads for visual grounding. In *CVPR*, pp. 9339–9350. Computer Vision Foundation / IEEE, 2025b.

Jaewoo Lee, Boyang Li, and Sung Ju Hwang. Concept-skill transferability-based data selection for large vision-language models. In *EMNLP*, pp. 5060–5080. Association for Computational Linguistics, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023a.

Shengzhi Li, Rongyu Lin, and Shichao Pei. Multi-modal preference alignment remedies degradation of visual instruction tuning on language models. In *ACL (1)*, pp. 14188–14200. Association for Computational Linguistics, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, pp. 292–305. Association for Computational Linguistics, 2023b.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: on pre-training for visual language models. In *CVPR*, pp. 26679–26689. IEEE, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pp. 26286–26296. IEEE, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *ECCV (6)*, volume 15064 of *Lecture Notes in Computer Science*, 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. In *ACM Multimedia*, pp. 3712–3721. ACM, 2025.

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. In *NeurIPS*, 2022.

Xudong Lu, Yinghao Chen, Renshou Wu, Haohao Gao, Xi Chen, Xue Yang, Xiangyu Zhao, Aojun Zhou, Fangyuan Li, Yafei Wen, Xiaoxin Chen, Shuai Ren, and Hongsheng Li. Genieblue: Integrating both linguistic and multimodal capabilities for large language models on mobile devices. In *ICCV*, pp. 4198–4210. IEEE, 2025.

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. *CoRR*, abs/2309.04564, 2023.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *EMNLP*, pp. 2381–2391. Association for Computational Linguistics, 2018.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *NeurIPS*, pp. 20596–20607, 2021.

Bardia Safaei, Faizan Siddiqui, Jiacong Xu, Vishal M. Patel, and Shao-Yuan Lo. Filter images first, generate instructions later: Pre-instruction data selection for visual instruction tuning. In *CVPR*, pp. 14247–14256. Computer Vision Foundation / IEEE, 2025.

Llama Team. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024.

Jianhong Tu, Zhuohao Ni, Nicholas Crispino, Zihao Yu, Michael Bendersky, Beliz Gunel, Ruoxi Jia, Xin Liu, Lingjuan Lyu, Dawn Song, et al. Mlan: Language-based instruction tuning preserves and transfers knowledge in multimodal language models. *arXiv preprint arXiv:2411.10557*, 2024.

Bin Wang, Chunyu Xie, Dawei Leng, and Yuhui Yin. IAA: inner-adaptor architecture empowers frozen large language model with multimodal capabilities. In *AAAI*, pp. 21035–21043. AAAI Press, 2025.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *NeurIPS*, 2024.

Lai Wei, Zihao Jiang, Weiran Huang, and Lichao Sun. Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4. *CoRR*, abs/2308.12067, 2023.

Xindi Wu, Mengzhou Xia, Rulin Shao, Zhiwei Deng, Pang Wei Koh, and Olga Russakovsky. ICONS: influence consensus for vision-language data selection. *CoRR*, abs/2501.00654, 2025.

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. In *ACL (Findings)*, volume ACL 2024 of *Findings of ACL*, pp. 15271–15342. Association for Computational Linguistics, 2024.

Xu Yang, Chen Liu, and Ying Wei. Data selection matters: Towards robust instruction tuning of large multimodal models. In *NeurIPS*, 2025.

Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *CVPR*, pp. 23219–23230. IEEE, 2024.

Qifan Yu, Zhebei Shen, Zhongqi Yue, Yang Wu, Bosheng Qin, Wenqiao Zhang, Yunfei Li, Juncheng Li, Siliang Tang, and Yueting Zhuang. Mastering collaborative multi-modal data selection: A focus on informativeness, uniqueness, and representativeness. In *ICCV*, pp. 155–165. IEEE, 2025a.

Yahan Yu, Duzhen Zhang, Yong Ren, Xuanle Zhao, Xiuyi Chen, and Chenhui Chu. Progressive lora for multimodal continual instruction tuning. In *ACL (Findings)*, volume ACL 2025 of *Findings of ACL*, pp. 2779–2796. Association for Computational Linguistics, 2025b.

Zeping Yu and Sophia Ananiadou. Locate-then-merge: Neuron-level parameter fusion for mitigating catastrophic forgetting in multimodal llms. In *EMNLP (Findings)*, pp. 7065–7078. Association for Computational Linguistics, 2025.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, pp. 9556–9567. IEEE, 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *ACL (1)*, pp. 4791–4800. Association for Computational Linguistics, 2019.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models. In *NAACL (Findings)*, volume NAACL 2025 of *Findings of ACL*, pp. 881–916. Association for Computational Linguistics, 2025.

APPENDIX

A ANALYSIS OF DESIGN CHOICES

A.1 WHY IS THE FIRST-LAYER ATTENTION SUFFICIENT?

PLUV measures the perturbation introduced by visual information to language representations by comparing text attention patterns under text-only and vision-language inputs. A natural question is whether attention from all transformer layers is required, or whether the first layer alone is sufficient to characterize such perturbations. To investigate this question, we randomly sample 18766 examples from the LLaVA-665K dataset and conduct a cross-layer consistency analysis.

For the n-th sample at layer l, we denote the text-to-text attention matrices under the vision-language and text-only conditions as $A^l_{n,\upsilon\tau}$ and $A^l_{n,\tau}$, respectively. We perform SVD on each attention matrix:

$$A^l_{n,j} = U^l_{n,j}\Sigma^l_{n,j}\left(V^l_{n,j}\right)^\top, j \in \{\upsilon\tau, \tau\}.$$

We use the left singular vector corresponding to the largest singular value as the dominant representation of the attention matrix: $u^l_{n,j} = U^l_{n,j}[:, 1]$. We then compute the cosine similarity between the dominant attention directions under the text-only and vision-language conditions:

$$sim^l_n = \cos\left(u^l_{n,\upsilon\tau}, u^l_{n,\tau}\right).$$

For a dataset containing N samples, the similarity scores at layer l are given by $S^l = \left\{sim^l_1, sim^l_2, \ldots, sim^l_N\right\}$. To analyze consistency across layers, we partition the similarity scores into ten intervals: $[0, 0.1), [0.1, 0.2), \ldots, [0.9, 1.0]$. For layer l and similarity interval b, we define the set of samples whose similarity scores fall within b as $S^l_b = \left\{n \mid sim^l_n \in b\right\}$. Taking the first layer as the reference, we define the sample retention rate of layer l within interval b as:

$$R^l_b = \frac{\left|S^1_b \cap S^l_b\right|}{|S^1_b|}.$$

R^l_b measures the proportion of samples that fall into similarity interval b at the first layer and remain in the same interval at layer l. A higher R^l_b indicates stronger cross-layer consistency in how visual information perturbs text attention patterns.

As shown in Figure 7, the similarity distributions exhibit broadly consistent patterns across different layers, suggesting that sample-level perturbation patterns do not change substantially with model depth. In particular, although a large proportion of samples are concentrated in the intermediate similarity ranges, a subset of samples consistently falls into high-similarity intervals, such as $[0.7, 0.8)$ and above, across different layers. This indicates that, for these samples, introducing visual information has only a limited effect on text attention patterns, and their textual representations remain highly consistent between the text-only and vision-language conditions. Furthermore, the sample retention rates reported in Table 5 show that samples assigned to a particular similarity interval at the first layer largely remain in the same interval at subsequent layers. This suggests that the relative degree of perturbation among samples is largely established at an early stage and remains relatively stable throughout the subsequent layers. Taken together, the similarity distributions in Figure 7 and the retention-rate results in Table 5 indicate that the first layer already provides a reliable characterization of sample-level attention perturbations under different input conditions. Therefore, using only the first-layer attention provides an effective approximation of the behavior observed in later layers while avoiding the additional computational overhead of extracting and processing attention matrices from all transformer layers.

A.2 WHY IS THE LEADING SINGULAR COMPONENT SUFFICIENT?

PLUV uses the leading singular component of an attention matrix to characterize its dominant structure. To examine whether the leading singular component alone preserves sufficient information, we again randomly sample 18766 examples from the LLaVA-665K instruction-tuning dataset and analyze the energy distribution across different singular components.

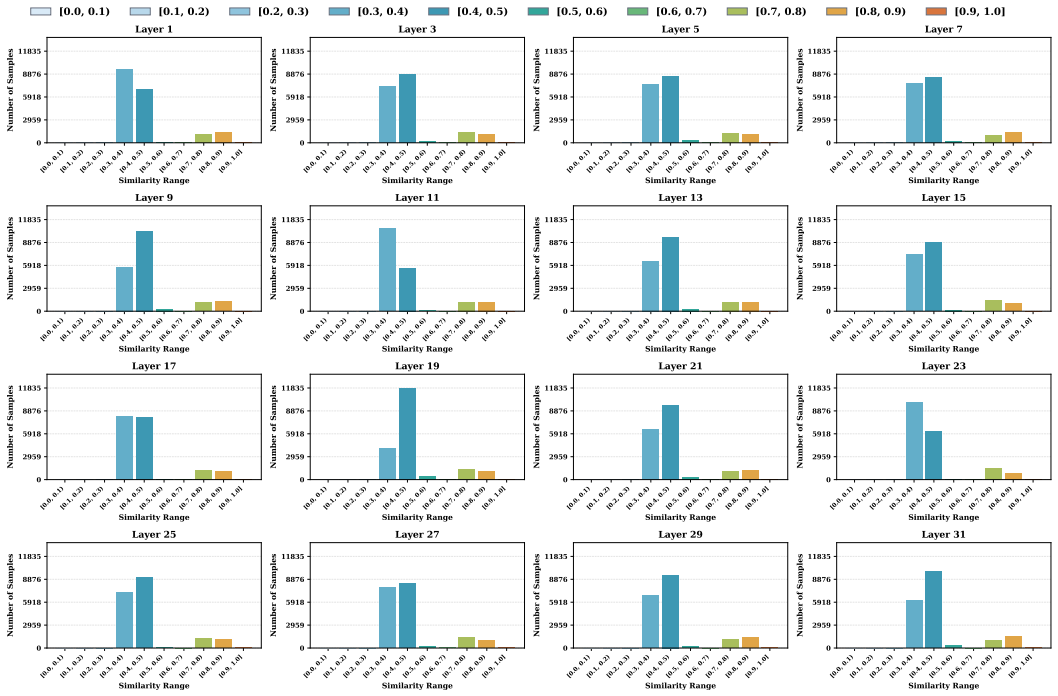


Figure 7: **Cross-layer distribution of text attention similarity between text-only and vision-language inputs.** Similarity is measured by the cosine similarity between the dominant singular components of the corresponding text-to-text attention matrices. The consistent distributions across layers indicate that sample-level perturbation patterns remain stable across the model depth.

Table 5: **Cross-layer sample retention rates across different similarity intervals.** The first layer is used as the reference, and the retention rate measures the proportion of samples that remain within the same similarity interval at subsequent layers. Higher retention rates indicate stronger cross-layer consistency in sample-level perturbation patterns.

Similarity Range	Samples	1	3	5	7	9	11	13	15	17	19	21	23	25	27	29	31
[0.0, 0.1)	0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
[0.1, 0.2)	0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
[0.2, 0.3)	0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
[0.3, 0.4)	9470	100.0	76.7	79.4	81.1	60.5	97.7	68.6	77.4	85.5	42.5	68.6	96.3	75.8	83.0	71.3	64.3
[0.4, 0.5)	6877	100.0	96.8	95.9	97.6	96.3	78.0	96.5	98.0	97.2	92.9	96.4	85.9	98.5	97.1	97.7	95.0
[0.5, 0.6)	17	100.0	100.0	100.0	94.1	94.1	58.8	100.0	88.2	76.5	100.0	100.0	82.4	100.0	94.1	100.0	100.0
[0.6, 0.7)	1	100.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0
[0.7, 0.8)	1084	100.0	95.8	92.8	82.7	91.3	91.5	93.4	96.1	91.6	96.7	89.6	88.9	95.3	97.4	88.4	79.8
[0.8, 0.9)	1305	100.0	77.5	80.5	93.3	86.8	84.9	86.2	76.0	78.5	76.8	87.2	60.5	83.7	72.8	90.6	94.0
[0.9, 1.0]	12	100.0	66.7	58.3	91.7	75.0	66.7	83.3	66.7	50.0	58.3	66.7	41.7	75.0	58.3	66.7	100.0

For the n-th sample, let its attention matrix be denoted by X_n. We perform SVD as $X_n = U_n \Sigma_n (V_n)^\top$, where $\Sigma_n = diag\left(\sigma_{n,1}, \sigma_{n,2}, \ldots\right)$, with $\sigma_{n,1} \geq \sigma_{n,2} \geq \cdots \geq 0$. We define the energy ratio of the k-th singular component as:

$$R_{n,k} = \frac{\sigma_{n,k}^2}{\sum_j \sigma_{n,j}^2}.$$

Where $R_{n,k}$ measures the contribution of the k-th singular component to the total Frobenius energy of the attention matrix. For the k-th singular component, its energy-ratio distribution over the entire dataset is $R_k = \{R_{1,k}, R_{2,k}, \ldots, R_{N,k}\}$. To further investigate how much information is retained by the first K singular components, we define the cumulative energy ratio as:

$$C_n(K) = \frac{\sum_{k=1}^{K} \sigma_{n,k}^2}{\sum_j \sigma_{n,j}^2}.$$

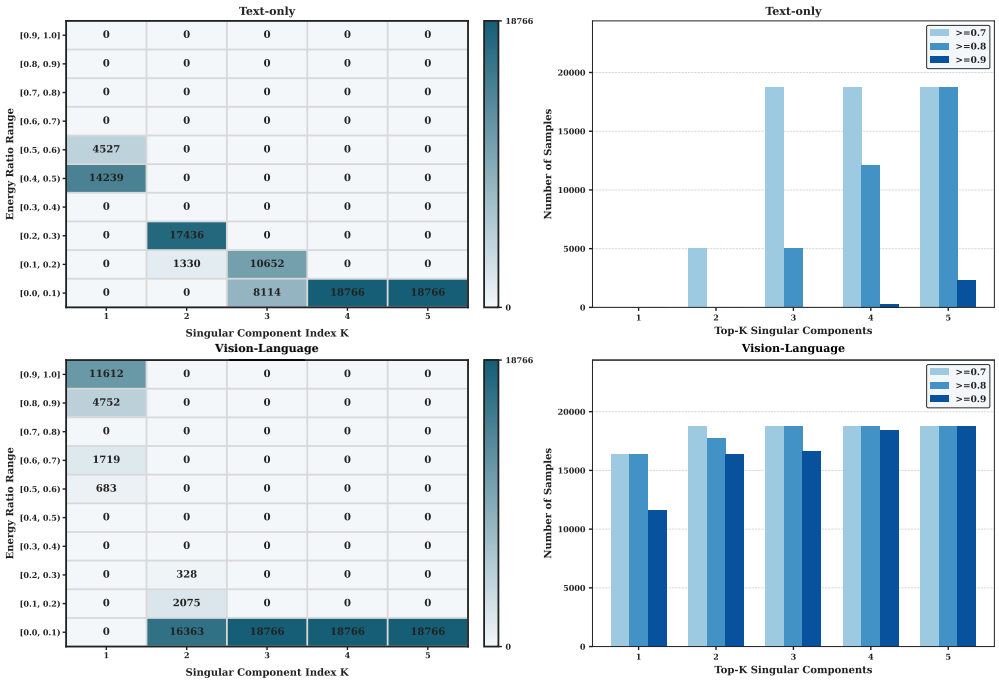


Figure 8: **Energy distributions of singular components under text-only and vision-language inputs.** The left panels show the number of samples across different energy-ratio ranges for each singular component index K, while the right panels report the number of samples whose cumulative energy reaches different thresholds as K increases.

For each K, we count the number of samples satisfying $C_n(K) \geq 0.7, C_n(K) \geq 0.8, C_n(K) \geq 0.9$, respectively, to quantify how effectively different numbers of leading singular components capture the energy of the attention matrix.

As shown on the left of Figure 8, under both the text-only and vision-language conditions, the singular-value energy of the attention matrices is strongly concentrated in the leading components. Specifically, the first few singular components account for substantially larger proportions of the total energy, while the contribution of subsequent components decreases rapidly as the component index increases. Furthermore, as shown in the right of Figure 8, the Top-K cumulative energy results indicate that as K increases, the number of samples reaching cumulative energy thresholds of 0.7, 0.8, and 0.9 steadily increases. This further demonstrates that most of the energy of the attention matrices is concentrated in a small number of leading singular components. Importantly, our goal is not to reconstruct the complete attention matrix, but rather to extract its dominant attention direction and compare how this direction changes between the two input conditions. Therefore, the leading singular component provides a compact representation of the dominant structure of the attention matrix, preserving the salient information while substantially reducing computational overhead.

B Detailed Experimental Setup

B.1 Implementation Details

All experiments are conducted on **two NVIDIA A6000 GPUs** with 48 GB of memory each to ensure a consistent hardware environment. For all visual instruction tuning experiments, we adopt the same training configuration and model initialization to ensure a fair comparison. All models are trained for **one epoch** following the official visual instruction tuning settings of their corresponding base models. Except for the data selection strategy, all other training hyperparameters are kept identical across different methods, ensuring that performance differences primarily arise from the selected training data.

B.2 Data Pools

We primarily conduct our experiments on the LLaVA-665K dataset. LLaVA-665K aggregates multiple vision-language instruction data sources and contains approximately 665K vision-language instruction samples, covering a diverse range of tasks, including visual question answering, image captioning, and complex visual reasoning. Each sample typically consists of an image, a textual instruction, and its corresponding response, providing rich vision-language supervision for VLMs. We use LLaVA-665K as our primary data selection pool and compare the effectiveness of different selection strategies on this pool.

To further evaluate the robustness of our method across different vision-language instruction data distributions, we additionally conduct experiments on VisionFlan-186K. VisionFlan-186K contains approximately 186K vision-language instruction samples spanning 191 distinct vision-language tasks. Compared with LLaVA-665K, VisionFlan-186K exhibits a more fine-grained task composition and a broader task distribution, making it suitable for evaluating the generalization ability of data selection baselines across different data sources and task distributions.

B.3 Evaluation Benchmarks

As shown in Table 6, we employ a suite of text-only and vision-language benchmarks to comprehensively evaluate the text-only and vision-language understanding capabilities of the models.

Text-only Benchmarks.

HellaSwag evaluates commonsense reasoning by requiring the model to select the most plausible continuation of a given scenario from multiple candidates. We report results on its validation split.

MMLU covers 57 subjects spanning stem, the humanities, social sciences, and other domains, and is designed to assess broad knowledge and reasoning capabilities. We evaluate on its test split.

MMLU-Pro is a more challenging extension of MMLU, featuring more difficult questions and an expanded set of answer choices to better assess complex knowledge reasoning. We use its test split for evaluation.

ARC evaluates grade-school science knowledge and reasoning and consists of two subsets, ARC-Easy (ARC-E) and ARC-Challenge (ARC-C). We evaluate both subsets on their test splits.

OpenBookQA consists of multiple-choice elementary science questions that require models to combine scientific facts with commonsense knowledge for reasoning. We report results on its test split.

Vision-Language Benchmarks.

ScienceQA evaluates scientific question answering and multi-step reasoning by requiring models to jointly reason over textual and visual information. We evaluate on its test split.

VizWiz contains images and questions collected from blind and visually impaired users in real-world scenarios, and evaluates visual question answering under challenging conditions such as low-quality or ambiguous images. We report results on its test split.

POPE is designed to evaluate object hallucination in vision-language models by examining whether a model incorrectly claims the presence of objects that do not exist in the image. We report results on its test split.

MMBench provides a comprehensive evaluation of visual perception and vision-language reasoning through multiple-choice questions. We use MMBench-Test EN for evaluation.

MME comprehensively evaluates vision-language models from both perception and cognition perspectives. We use its test set and separately report results for MME-Perception (MME-P) and MME-Cognition (MME-C).

MMMU targets college- and expert-level multidisciplinary knowledge and evaluates professional knowledge understanding and reasoning in complex vision-language scenarios. We report results on its validation split.

Table 6: Evaluation benchmarks and data splits

Benchmarks	Evaluation Focus	Splits
Text-only Benchmarks		
HellaSwag	Commonsense reasoning and sentence completion	Validation
MMLU	Knowledge and reasoning across 57 subjects	Test
MMLU-Pro	Challenging multi-domain knowledge and reasoning	Test
ARC-E / ARC-C	Grade-school science knowledge and reasoning	Test
OpenBookQA	Scientific and commonsense reasoning	Test
Vision-Language Benchmarks		
ScienceQA	Vision-language science question answering and reasoning	Test
VizWiz	Visual question answering in real-world scenarios	Test
POPE	Object hallucination evaluation	Test
MMBench	Visual perception and vision-language reasoning	Test (EN)
MME	Perception and cognition capabilities	Test
MMMU	Expert-level multidisciplinary vision-language reasoning	Validation

Table 7: Performance comparison on text-only benchmarks under different data selection strategies on VisionFlan-186K.

Methods	HellaSwag ↑	MMLU ↑	MMLU-PRO ↑	ARC-E ↑	ARC-C ↑	O.B.QA ↑	Rel (%)
Full-Data (186K)	70.8	45.0	12.2	75.1	45.1	45.0	100%
Random (57K)	73.5	46.0	10.1	76.6	46.4	45.3	99.1%
PRISM (57K)	74.3	44.5	10.7	76.2	46.8	46.0	99.8%
PLUV (28K)	**73.7**	**44.7**	**12.4**	**78.8**	**46.8**	**44.6**	**102.1%**

C Generalization to VisionFlan-186K

To further evaluate whether PLUV can generalize beyond the original visual instruction data pool, we conduct additional experiments on VisionFlan-186K. We adopt the same training and evaluation settings as in the main experiments and compare PLUV with full-data training, random selection, and PRISM. The results on text-only and vision-language benchmarks are reported in Tables 7 and 8, respectively.

C.1 Text-only Performance

As shown in Table 7, despite using only 28K samples, PLUV maintains language performance on text-only benchmarks, achieving an overall relative performance of 102.1%. This demonstrates that, after vision-language instruction tuning on the VisionFlan-186K data pool, PLUV preserves the model's original language capability more effectively than other data selection baselines. Furthermore, these results indicate that even when the vision-language instruction data pool is changed, PLUV can consistently identify training samples with low language forgetting risk, thereby effectively mitigating the degradation of language capability during vision-language instruction tuning.

C.2 Vision-Language Performance

As shown in Table 8, PLUV also maintains vision-language capabilities on VisionFlan-186K. Using only 28K selected samples, PLUV achieves an overall relative performance of 101.2%, surpassing the full-data baseline while substantially reducing the number of training samples. Specifically, PLUV outperforms full-data training on SQA-I, MMBench, and MME, while maintaining comparable performance on the remaining benchmarks. These results further demonstrate that preserving language capabilities does not come at the expense of vision-language learning.

Table 8: Performance comparison on vision-language benchmarks under different data selection strategies on VisionFlan-186K.

Methods	SQA-I ↑	VizWiz ↑	POPE ↑	MMBench ↑	MME ↑	TextVQA ↑	Rel. (%)
Full-Data (186K)	60.8	41.7	83.4	52.6	1263.2	50.4	100.0%
Random (57K)	56.5	38.8	83.1	48.9	1175.0	46.9	94.1%
PRISM (57K)	61.1	42.3	84.1	53.1	1275.5	50.8	100.9%
PLUV(28K)	**62.1**	**41.4**	**83.4**	**53.4**	**1367.1**	**48.5**	**101.2%**

D Detailed analysis of data selection cost

As described in Sections 3.1 and 3.2, PLUV introduces additional computational overhead during the data selection stage, primarily arising from the two data analysis stages, LFRE and VDE. In the following, we provide a more detailed breakdown of the computational costs associated with these two stages and further discuss the practical advantages of PLUV.

D.1 Cost of LFRE

LFRE estimates the potential language forgetting risk of each sample by comparing the internal attention structures under two input views: the vision-language view and its corresponding text-only view. Therefore, each sample needs to be processed under both input conditions. For each view, LFRE extracts the text-to-text attention matrix from the first transformer layer of the frozen language model and averages the matrices across all attention heads. SVD is then applied to the resulting attention matrix to obtain its dominant singular component, and the cosine similarity between the dominant components of the two views is used as the LFRE score. Although LFRE only uses the first transformer layer and the dominant singular component, constructing and processing two input views for each sample still introduces additional computation. Moreover, the extraction of attention matrices and the subsequent SVD analysis also incur additional overhead. As shown in Table 9, LFRE requires 25 hours to process 665298 samples, with a total computational cost of 8.52×10^{18} FLOPs and a peak GPU memory usage of 41.8 GB. Therefore, in our experiments, LFRE constitutes the largest portion of the runtime of the PLUV data selection pipeline.

D.2 Cost of VDE

VDE further assesses the visual learning value of vision-language instruction samples by estimating their dependency on visual information. Specifically, VDE applies random joint masks to image-to-text attention heads and records the changes in model scores under different perturbation configurations. These score changes are subsequently used in ridge regression to estimate the contribution of each attention head to visual information, from which a sample-level visual dependency score is derived. As shown in Table 9, visual dependency estimation is only applicable to samples containing visual inputs. Therefore, text-only samples in the data pool do not need to be processed by VDE, and VDE evaluates 624608 vision-language samples in total. The overall runtime of VDE is 16 hours, with a computational cost of 8.01×10^{18} FLOPs and a peak GPU memory usage of 35.9 GB. These results show that, despite performing fine-grained attention-head-level analysis to estimate visual dependency, the overall computational overhead of VDE remains manageable.

D.3 Advantages of PLUV

The additional computation introduced by PLUV occurs only during the one-time, training-free data selection stage. Both LFRE and VDE are performed with frozen model parameters and require no backpropagation, gradient computation, or parameter updates. Therefore, PLUV does not introduce any additional model training process. Moreover, PLUV adopts a coarse-to-fine two-stage data selection strategy. In the practical data selection pipeline, LFRE first identifies a candidate subset $\mathcal{D}_{LFRE}$ with relatively low language forgetting risk from the full data pool. VDE is then performed only on this candidate subset, rather than further analyzing all vision-language samples in the entire data pool. By progressively reducing the search space across stages, this design further decreases the practical selection cost of the complete PLUV pipeline.

Table 9: Computational cost breakdown of PLUV.

Module	Evaluated Samples	Hours	FLOPs	Peak GPU Memory (GB)
LFRE	665298	25	8.52×10^{18}	41.8
VDE	624608	16	8.01×10^{18}	35.9

Table 10: Text-only performance under different similarity thresholds and retained ratios.

Similarity Threshold	$\lvert\mathcal{D}_{LFRE}\rvert$	Retained Ratio	$\lvert\mathcal{D}_{PLUV}\rvert$	HellaSwag ↑	MMLU ↑	MMLU-PRO ↑	ARC-E ↑	ARC-C ↑	O.B.QA ↑	Rel. (%)
Full-Finetune	-	-	-	75.7	50.5	15.9	78.8	49.7	46.8	100%
0.8	78024	0.3	51889	75.2	49.7	13.7	78.5	50.0	46.2	97.1%
		0.4	55623	75.7	50.0	14.8	78.5	50.3	46.8	98.8%
		0.5	59356	75.3	49.5	13.5	78.4	50.6	45.6	96.9%
0.7	119287	0.3	64268	75.6	49.9	11.7	78.7	50.1	44.6	94.7%
		0.4	72128	75.8	50.0	17.3	79.6	50.3	45.4	101.2%
		0.5	79988	75.4	50.0	13.8	78.9	49.9	46.6	97.6%
0.4	471487	0.3	169928	75.6	49.4	12.9	78.8	49.2	47.6	96.6%
		0.4	213007	75.3	50.3	12.0	78.6	50.3	46.6	95.8%
		0.5	256087	75.4	50.5	11.9	79.1	48.9	47.2	95.7%

Table 11: Vision-language performance under different similarity thresholds and retained ratios.

Similarity Threshold	$\lvert\mathcal{D}_{LFRE}\rvert$	Retained Ratio	$\lvert\mathcal{D}_{PLUV}\rvert$	SQA ↑	SQA-I ↑	VizWiz ↑	POPE ↑	MMBench ↑	MME-P ↑	MME-C ↑	MMMU ↑	Rel. (%)
Full-Finetune	-	-	-	69.4	66.8	50.0	85.9	64.3	1510.7	311.9	35.4	100%
0.8	78024	0.3	51889	68.4	67.9	53.5	86.3	62.5	1385.4	327.1	35.4	100.2%
		0.4	55623	68.7	67.5	53.1	87.5	63.3	1365.8	311.4	35.7	99.7%
		0.5	59356	68.8	67.7	54.1	86.4	62.6	1394.5	250.7	37.1	98.0%
0.7	119287	0.3	64268	69.6	67.5	52.6	86.0	64.0	1377.0	344.6	34.2	100.6%
		0.4	72128	70.3	69.3	52.0	87.3	65.4	1432.8	328.2	34.8	101.3%
		0.5	79988	69.5	67.9	50.7	87.2	63.9	1482.1	289.3	35.1	99.3%
0.4	471487	0.3	169928	70.0	68.4	51.5	86.2	64.3	1477.6	382.5	35.8	103.5%
		0.4	213007	70.0	68.4	50.0	86.4	66.1	1509.6	330.7	35.1	101.5%
		0.5	256087	70.3	68.9	53.4	86.7	66.3	1504.2	349.6	36.1	103.6%

In practice, both LFRE and VDE support efficient batch processing. For LFRE, the vision-language views and text-only views of multiple samples can be separately organized into batches for model forward passes, thereby improving GPU utilization. For VDE, multiple samples and their corresponding attention-head perturbation evaluations can likewise be processed in batches. In addition, the attention matrices extracted during LFRE can be cached and reused, avoiding redundant model forward passes and repeated attention analysis. These properties provide PLUV with strong parallelizability and scalability, allowing the data selection cost to remain within a reasonable range without requiring any modification to the subsequent vision-language instruction tuning pipeline.

E Detailed performance for selection threshold and retained ratio

To comprehensively analyze the effects of the similarity threshold in LFRE and the retained ratio in VDE, we report detailed experimental results under different selection configurations. As shown in Tables 10 and 11, we evaluate various combinations of similarity thresholds and retained ratios on both text-only and vision-language benchmarks. Here, $|\mathcal{D}_{LFRE}|$ and $|\mathcal{D}_{PLUV}|$ denote the numbers of samples in the candidate subset obtained after LFRE and the final selected subset obtained after VDE, respectively. Different selection configurations lead to variations in both the final subset size and model performance. By analyzing performance variations under different configurations, we further investigate how the two-stage data selection process affects the model's text-only and vision-language capabilities. Among these configurations, the setting with a similarity threshold of 0.7 and a retained ratio of 0.4 achieves a favorable balance between the number of selected samples and performance, further validating the effectiveness of the two-stage data selection strategy.

F THE USE OF LARGE LANGUAGE MODELS

During the preparation of this paper, the use of Large Language Models was strictly limited to language polishing and text translation. All core scientific contributions—including research conceptualization, experimental design and execution, data analysis, and conclusion derivation—were independently completed by the authors, who assume full responsibility for this work. Following a rigorous verification, the final text has been ensured to be accurate, factually correct, and in full compliance with academic integrity standards.