Train a Unified Multimodal Data Quality Classifier with Synthetic Data

Weizhi Wang[1,2] **Rongmei Lin**[2] **Shiyang Li**[2] **Colin Lockard**[2] **Ritesh Sarkhel**[2]
Sanket Lokegaonkar[2] **Jingbo Shang**[2,3] **Xifeng Yan**[1] **Nasser Zalmout**[2] **Xian Li**[2]
[1]UC Santa Barbara [2]Amazon Stores Foundational AI [3]UC San Diego
https://victorwz.github.io/UniFilter/

Abstract

The Multimodal Large Language Models (MLLMs) are continually pre-trained on a mixture of image-text caption data and interleaved document data, while the high-quality data filtering towards image-text interleaved document data is under-explored. We propose to train an efficient MLLM as a Unified Mulitmodal Data Quality Classifier to Filter both high-quality image-text caption and interleaved data (UniFilter). To address the challenge of collecting diverse labeled multimodal data, we introduce a semi-synthetic approach that leverages readily available raw images and generates corresponding text across four quality levels. This method enables efficient creation of sample-score pairs for both caption and interleaved document data to train UniFilter. We apply UniFilter to curate high-quality caption data from DataComp caption dataset and interleaved data from the OBELICS image-text interleaved dataset. MLLMs pre-trained on the filtered data demonstrate significantly enhanced capabilities compared to those trained on baseline-filtered data, achieving stronger zero-shot reasoning and in-context learning capabilities. After visual supervised fine-tuning, these UniFilter-induced MLLMs achieve stronger performance on various benchmarks, highlighting the downstream benefits of high-quality multimodal pre-training.

1 Introduction

Large-scale multimodal datasets significantly motivates the recent advances in Vision Language Models (VLMs) (Radford et al., 2021; Li et al., 2022; Wang et al., 2021) and Multimodal Large Language Models (MLLMs) (Lin et al., 2024; McKinzie et al., 2024; Laurençon et al., 2024b; Xue et al., 2024). The scaled up data allows the MLLMs to harvest the knowledge in the training corpora to the greatest extent and promotes the state-of-the-art MLLMs. The MLLMs are trained on a mixture of image-text caption data and interleaved document data to enhance both zero-shot and few-shot capability. Moreover, with the limited computing resources but the overwhelming number of data mining from CommonCrawl Snapshots, the recent large-scale MLLMs are only trained on a data subset for less than one epoch. Therefore, the data quality became the major bottleneck in training stronger models. In selecting high-quality image-text caption dataset, the representative model-based filter, CLIPScore filter (Schuhmann et al., 2021; Gadre et al., 2023) has become the predominant data filtering method. However, CLIPScore can only deal with image-text caption data based on the similarity between a single image and a short text caption. It is completely un-explored on how to select high-quality image-text interleaved data, which contains multiple images and long text paragraphs interleaving in one document.

To address this problem, we propose to train an efficient MLLM as a **Uni**fied Mulitmodal Data Quality Classifier to **Filter** both high-quality image-text caption and interleaved data (**UniFilter**). Adopting an MLLM architecture for the proposed data quality classifier effectively overcomes the limitation of CLIPScore, which can only process single image-text pairs. The proposed UniFilter can process both image-text paired and interleaved data and output a float quality score to indicate the quality of this multimodal data sample. Meanwhile, it outperforms CLIPScore on curating high-quality image-text caption data for enhancing both VLM and MLLM pre-training. Simultaneously, it achieves a high inference throughput of 130 samples/s by leveraging Qwen-2.5-0.5b as LLM backbone, sligtly outperforming the CLIPScore method's 128 samples/s on the same hardware.

The key to train an effective data quality classifier lies in constructing accurate sample-score pairs (Dubey et al., 2024; Penedo et al., 2024). Human annotations for these pairs are costly and challenging to maintain consistency across different

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 1972–1986
November 4-9, 2025

annotators. To address this, we propose a novel semi-synthetic multimodal data generation method by leveraging the proprietary MLLMs. Given that the proprietary MLLMs excel in text generation given multimodal inputs and the raw images are readily available, we sample a diverse set of original images from captioned or interleaved data. We then use proprietary MLLMs to generate the full multimodal data following quality requirements across 4 quality levels (Section 2.1), in which a similar 4 level quality score is also used in FineWeb-Edu-Quality-Classifier (Penedo et al., 2024). Then the synthetic data can be easily constructed as sample-score pairs, with score labels 0, 1, 2, and 3 corresponding to the defined quality levels in the prompts.

In addition to well-designed synthetic training data construction, we conduct comprehensive ablation studies on the effective and efficient multimodal model architecture of UniFilter on the held-out validation synthetic sample-score data. We experiment with 6 combinations of choices of the vision encoder, visual projector, and the LLM backbone for constructing the UniFilter architecture. The architecture designs of SigLIP-SO-400M vision encoder (Zhai et al., 2023), adaptive average pooling projector and the Qwen-2.5-0.5B LLM (Yang et al., 2024) achieves the best trade-off between data quality classification performance and efficiency.

We conduct comprehensive experiments on image-text caption data and interleaved document data filtering to demonstrate the effectiveness of our method over strong baselines. We firstly validate the priority of UniFilter on curating high-quality image-text caption data over strong baselines through the experiments on MLLM pre-training with curated caption data only. Our method UniFilter outperforms the state-of-the-art (SOTA) CLIP-based filtering method, Data Filtering Network (DFN) (Fang et al., 2023) and the SOTAQ MLLM-based data filtering model, MLM-Filter (Wang et al., 2024) on all 5 zero-shot VQA datasets. Secondly, pre-training MLLMs on the high-quality image-text interleaved document data curated by UniFilter promotes the few-shot learning capabilities of MLLMs by +0.7 and +2.8 average scores on 4-shot and 8-shot VQA performance over the baseline data filters. Finally, instruction-tuned MLLM based on UniFilter pre-training outperforms baselines and achieves +3.1 average improvement on VQA tasks and +1.5 improvement on MMMU benchmark, demonstrating the broad benefits of our approach. The summarization of the contributions of UniFilter are as follows:

1. We introduce UniFilter, the first unified approach for filtering both image-text caption and interleaved document data. By leveraging an MLLM-based architecture, UniFilter overcomes the limitations of existing methods that can only process single image-text pairs, enabling effective quality assessment of complex multimodal data structures.

2. We propose an efficient semi-synthetic data generation method that combines original images with synthetic text across multiple quality levels. This approach addresses the challenge of obtaining diverse, labeled multimodal data for classifier training, enabling scalable and cost-effective creation of high-quality training datasets.

3. MLLMs pre-trained on high-quality caption data and interleaved document data filtered by UniFilter demonstrate significant performance improvements over models trained with baseline data filtering methods. These gains from high-quality pre-training also persist after the SFT stage, further enhancing model capabilities.

2 Synthetic Data Construction

Compared with collecting training data of data quality classification tasks from web-resources or human annotators, the synthetic data can be easily generated on a large scale to provide sufficient training data for models, which is a more feasible solution to empowering the effective training of data quality classifier. Furthermore, by incorporating controlled 4-level quality requirements (Table 1) into the data generation prompts, synthetic data generation can adhere strictly to these designated quality standards, ensuring clear quality boundaries among data at different quality levels. This proposed approach guarantees data sufficiency of quality classification tasks for effective training of data quality classifier and meanwhile enhances the generalization capabilities of the classifiers on data across multiple quality levels. The synthetic data generation pipeline is shown in Figure 1.

2.1 Define Data Quality Requirements

We firstly design a fine-grained data quality taxonomy on the multimodal data. Instead of using a

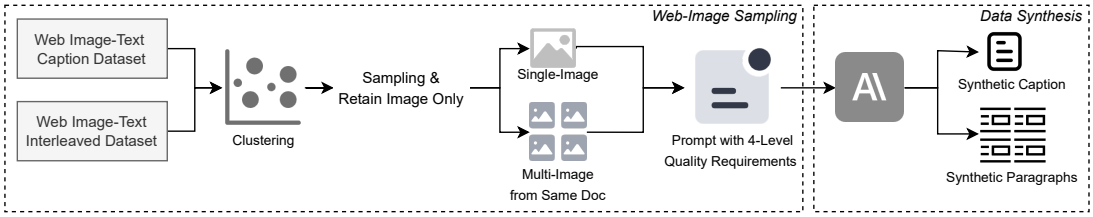

Figure 1: The pipeline of semi-synthetic data generation for image-text caption data and interleaved document data.

binary classification of positive and negative, we establish four quality levels: easy negative, medium negative, hard negative, and positive. These four quality levels are designed to capture the spectrum of data quality typically encountered in real-world multimodal datasets. The "easy negative" category represents completely irrelevant or nonsensical data, while "medium negative" captures data with significant but not entirely unrelated errors. "Hard negative" simulates subtle mismatches or minor inaccuracies that are challenging to detect, and "positive" represents high-quality, well-aligned multimodal data.

This granular approach enables our classifier to learn discriminative features across a range of quality levels, enhancing its ability to filter real-world data effectively. We develop different prompts for both caption and interleaved data to accurately describe each quality level, shown in Table 1. These prompts guide Claude-3-Sonnet (Anthropic, 2024) in generating synthetic multimodal data that follows the specified quality requirements. The full prompt giving to Claude-3-Sonnet are presented in Appendix E.

2.2 Semi-Synthetic Data Generation

The synthetic data on the proposed mutlimodal data quality classification tasks should be diversified and generalized on both image and text sides to train a generalized multimodal data quality classifier. We initially considered generating fully synthetic multimodal data. However, we found that the SOTA image generation models like Midjourney and Dalle-3 (Betker et al., 2023) are stuck into specific image styles, i.e. carton, due to the post-training adaptations for these diffusion models. Thus, we adopted a semi-synthetic approach: sampling original images from web-crawled caption and interleaved document datasets, while using Claude-3-Sonnet (Anthropic, 2024) to generate corresponding text at various quality levels.

This semi-synthetic approach offers several advantages. It ensures visual diversity and realism by using real-world images while allowing for controlled text generation at various quality levels. This method is highly scalable and efficient, enabling the creation of large, diverse datasets for classifier training.

For our semi-synthetic data generation, we selected two prominent datasets as our source data: DataComp for image-text captions and OBELICS for interleaved documents. To enhance the diversity and topic coverage on the image sampling process, we cluster the DataComp-small images into 10k clusters based on their image embeddings extracted by CLIP ViT-L/14. Then we select 4 images from each cluster to get the original 40k image data for the synthetic caption data generation. For the interleaved document data from OBELICS, we compute the average pooling of image embeddings of all images within a single document to create a representative visual embedding for each document. We then clustered the OBELICS dataset into 10k document-level clusters and sampled 40k image-text interleaved documents from these clusters.

Finally, we generate 40k synthetic caption data and 40k synthetic interleaved document data across 4 designed quality levels. We assign the integer quality scores to each synthetic sample while the quality levels of easy negative, medium negative, hard negative, and positive are corresponding to 0, 1, 2, and 3. We held 5% of 80k data as the validation set for further model developments on the proposed multimodal data quality classification tasks. After collecting original synthetic data, we adopt `Llama-guard-3-8B` (Dubey et al., 2024) to efficiently scan the synthetic text and ensure there is no safety concerns in the generated texts. We also include 4k non-synthetic high-quality image-caption data from MSCOCO (Lin et al., 2014) and Flickr (Young et al., 2014) into the final dataset, of which these 4k non-synthetic data are all assigned with quality score of positive. We use the joint sample-score paired data of both image-text caption and interleaved data to train a SINGLE unified multimodal data quality classifier, which can pro-

Quality Level	Quality Requirements in Prompt
Easy Negative	a negative image caption which is completely unrelated to this image.
Medium Negative	a negative image caption which has remarkable errors in describing the image.
Hard Negative	a hard negative image caption which has subtle difference with the positive caption. The negative caption contains only one property error in describing the image.
Positive	a high-quality, comprehensive, detail-enriched caption for this image.
Easy Negative	This document should involve many errors in writing and the document itself is not fluent in reading. The images and the text in the document should be completely not related. The images are inserted in inappropriate and arbitrary places in the document. This document should be knowledge limited and has no educational value to be used as textbooks in primary school or grade school teaching.
Medium Negative	This document is readable but still contains several writing errors. The images and document text are under the same topic and the text contents are still not aligned well to the images. The document is knowledge sparse and has very limited educational value to be used as textbooks in primary school or grade school teaching.
Hard Negative	This document should involve several errors in writing. The images and the text in the document are partially related. However, the images cannot help the understanding of the text and cannot provide any additional information. The images are inserted in reasonable places in the document. This document should contain several factual or commonsense knowledge errors which makes it inappropriate for educational purposes.
Positive	This document is a high-quality, comprehensive, detail-enriched document. The images are inserted in the appropriate places in the document to provide additional information to the statement or provide the background information.

Table 1: Data quality requirements for synthetic caption data and interleaved document data generation.

cess both image-text caption data and interleaved document data.

3 UniFilter Architecture

To achieve a unified architecture to process both the image-text caption and interleaved data, we construct the UniFilter based on a MLLM architecture. Figure 2 presents how a MLLM-based multimodal data quality classifier can process the two major types of image-text data. For the image-text interleaved data, the images and texts are encoded separately with vision encoder and word embedding layer and then reconstructed in original interleaving order. For the caption data, the image encoding and caption embeddings are concatenated and forwarded into the LLM backbone. A trainable one-dimensional classification head is appended on the top of LLM backbone to output a logit indicating the quality score of the input caption or interleaved data sample.

Adopting a MLLM-based data quality classifier can substantially improve the quality classification performance compared with CLIP-based architectures, while the introduced billion-level model parameters will bring huge inference cost to the data quality label inferences on the pre-training data scale. In order to to train the UniFilter to be both efficient and capable, we perform comprehensive ablation study on the MLLM architecture of the UniFilter. The recent advances on the architecture design of MLLMs (Liu et al., 2023b; Chen et al., 2023b; Bai et al., 2023) all deploy the modality fusion architecture with 3 major modules of vision encoder, vision-language projector, and the LLM. We inherit this architecture and perform detailed ablations on design choices on different modules. The configuration for model architecture ablations om each module is as follows:

- Vision Encoder: We choose two models with different input image resolutions of CLIP-ViT-Large-224px and CLIP-ViT-Large-336px from CLIP model family as well as SigLIP-ViT-SO400m-384px (Zhai et al., 2023) for the ablation studies on vision encoders.

- Visual Projector: We consider two type of projector architecture of the non-compressive Multi-Layer Perceptron (MLP) with 2x inner embedding size used in LLaVA (Liu et al., 2023b), and the compressive two-dimensional Adaptive Average Pooling layer with MLP used in DECO (Yao et al., 2024).

- LLM Backbone: We experiment with four representative small LLMs—Phi-3-mini-3.8B, Gemma-2-2B, and Qwen-2.5 (1.5B and 0.5B), as the base LLM for UniFilter. Larger LLMs

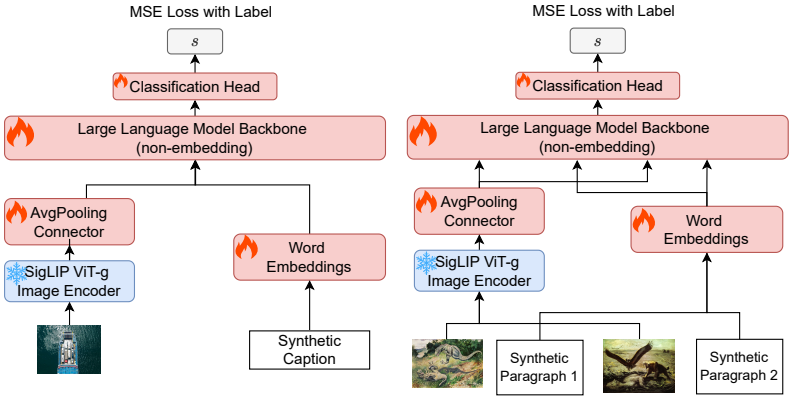

Figure 2: The unified model architecture of UniFilter which uses an efficient MLLM to classify the quality scores of both image-text paired data (*Left*) and interleaved data (*Right*).

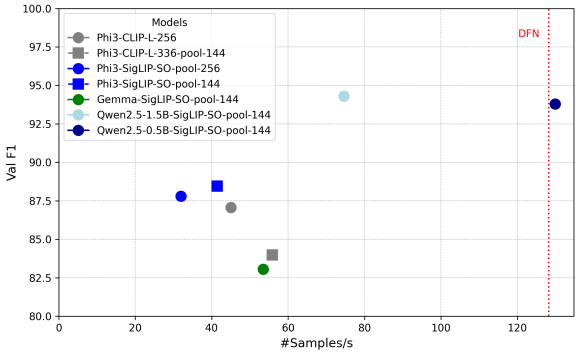

Figure 3: The classification F1 versus inference speed of different MLLM architecture ablation configurations on the held validation data of quality classification task.

with more than 4B parameters exceeds efficiency requirements for generating quality scores on pre-training data scale. The pre-trained language modeling head (Embd_Size$\times$Vocab_Size) is deprecated while a newly-initialized classification head (Embd_Size $\times$ 1) is trained for UniFilter. Then the output scalar logit is aligned to the synthetic quality label using Mean-Square-Error (MSE) loss.

Because of the quadratic time complexity of LLMs with respect to number of concatenated multimodal input tokens, the computation efficiency of MLLMs are heavily affected by the number of image tokens for representing one image. Therefore, conducting compression on image patches to fixed number of image tokens is mandatory to ensure the efficiency, especially for high-resolution vision encoders, i.e. SigLIP-so400m-384px and CLIP-Large-336px. Yao et al. (2024) compares the performance of 4 popular compressive vision projectors, Q-Former (Li et al., 2023a), C-Abstractor (Cha et al., 2024), D-Abstractor (Cha et al., 2024) and AdaptiveAveragePooling (AvgPool), and the AvgPool significantly outperforms other competitors. Thus, we adopt the two-dimensional AvgPool as the compressive vision projector and compare it with the non-compressive MLP projector.

We train each UniFilter variant for 10 epochs on synthetic contrastive data and the best model is selected based on validation accuracy. The UniFilter based on Qwen2.5-1.5b achieves the best quality classification performance while introducing significant computational overhead due to the additional 1b parameters compared with Qwen2.5-0.5b model. Among all group of MLLM architecture configurations, the architecture with SigLIP-SO400M vision encoder, AvgPool visual projector of 144 tokens per image, and the Qwen-2.5-0.5B LLM achieves the best trade-off between quality classification performance and efficiency in Figure 3, which is used as the final UniFilter model. Surprisingly, the final UniFilter model can achieve comparable inference speed with DFN-CLIP-Large (Fang et al., 2023).

4 Experiments

4.1 MLLM Pre-Training on Image-Text Caption Data Only

We apply each filtering method, including UniFilter, to curate high-quality image-text caption data from the DataComp-medium-128M pool (Gadre et al., 2023), and subsequently pre-train separate MLLMs using the datasets curated by the respective filtering methods. The DataComp-medium-128M pool is a noisy, web-crawled image-text caption dataset that employs only basic rule-based filtering, which is specifically designed to evaluate the effectiveness of model-based filtering

LLM	Vision Encoder	Projector	#Tokens per Image	Image Resolution	Validation Acc	Validation F1
Phi-3-3.8b	CLIP-L	MLP	256	224	90.8	87.1
Phi-3-3.8b	CLIP-L	AvgPool+MLP	144	336	88.7	84.0
Phi-3-3.8b	SigLIP-SO-400M	AvgPool+MLP	256	384	91.5	87.8
Phi-3-3.8b	SigLIP-SO-400M	AvgPool+MLP	144	384	91.9	88.5
Gemma-2-2b	SigLIP-SO-400M	AvgPool+MLP	144	384	88.2	83.1
Qwen2.5-1.5b	SigLIP-SO-400M	AvgPool+MLP	144	384	**95.2**	**94.3**
Qwen2.5-0.5b	SigLIP-SO-400M	AvgPool+MLP	144	384	94.8	93.8

Table 2: Ablation studies on the MLLM architecture of UniFilter.

Methods	GQA	VQA-v2	VizWiz	OKVQA	TextVQA	Avg.
DFN (Fang et al., 2023)	25.8	39.6	21.6	26.0	30.7	28.7
MLMFilter-Image-Text-Matching (Wang et al., 2024)	28.3	42.7	21.7	26.0	31.6	30.2
MLMFilter-Object-Detail-Fulfillment (Wang et al., 2024)	28.1	39.1	20.4	27.7	31.6	29.4
MLMFilter-Caption-Text-Quality (Wang et al., 2024)	28.4	40.7	20.3	27.2	**35.2**	30.4
MLMFilter-Semantic-Understanding (Wang et al., 2024)	24.0	40.8	18.5	23.8	29.8	27.4
UniFilter	**29.6**	**43.2**	**22.9**	**28.2**	32.5	**31.3**

Table 3: Zero-shot multimodal benchmark results of different pre-trained base MLLMs which are trained on only curated caption data for 5B tokens.

methods in selecting high-quality caption data.

Baselines We pick the following baseline methods for fair comparisons on MLLM pre-training: 1) Data-Filtering-Network (DFN), a strong CLIP-based data filtering model, which continually pre-trains the OpenAI CLIP-large on high-quality caption dataset for data filtering purpose; 2) MLM-Filter (Wang et al., 2024), which fine-tunes a MLLM to generate quality scores for caption data filtering with 4 different scoring metrics, Image-Text Matching (ITM), Object Detail Fulfillment (ODF), Caption Text Quality (CTQ), and Semantic Understanding (SU).

Training Setup. We compare the baseline and our method using the same MLLM architecture and training settings. The model architecture we adopt in MLLM pre-training consists of 3 modules of SigLIP-so400m vision encoder, AvgPool visual projector with 144 tokens per image, and the Phi-3-mini-3.8b LLM. The vision encoder is frozen at all time while other parameters are trainable. To ensure the fair comparisons, we set a fixed **30%** fraction of retained high-quality subset from DataComp-medium-128M pool for each filtering method, which can be tokenized into about 6B multimodal tokens for pre-training. Then, each MLLM is trained on filtered image-text caption data by each filtering method for **5B** multimodal tokens, eliminating the effects of slightly different number of tokens in training for one epoch for each filtered dataset. Other hyper-parameters and details for multimodal pre-training are presented in Appendix A. Additionally, we perform an ablation study for the filtering fraction hyperparamter in Appendix G.

Evaluation Benchmarks. We evaluate the zero-shot performance of each **base** pre-trained MLLMs on 5 visual-question answering datasets, including GQA (Hudson and Manning, 2019), VQA-v2 (Goyal et al., 2017), VizWiz (Gurari et al., 2018), TextVQA (Singh et al., 2019), and OKVQA (Marino et al., 2019). Among the 5 VQA datasets, the VQA-v2 and OKVQA focus on the commonsense knowledge understanding in the images. GQA and VizWiz emphasizes on the scene and spatial understandings and TextVQA lies on evaluating the OCR capability.

Results. The results in Table 3 demonstrate the superiority of UniFilter on curating image-text caption data for enhancing the understanding and reasoning capabilities of base non-sft base MLLMs. Moreover, the base MLLM pre-trained UniFilter curated caption data outperforms both DFN and all MLM-Filter metrics on the average performance of 5 multimodal benchmarks, demonstrating the strong generalization and diversity of the UniFilter-curated caption data in enhancing the capabilities of MLLMs across OCR, general reasoning, knowledge-reasoning, and scene reasoning. The MLLM trained with UniFilter curated data only lags behind the MLM-CTQ metric on TextVQA task, in which TextVQA dataset requires models to read and reason about text in images. MLMFilter-CTQ metric can effctively differentiate the caption

data with great text quality and might be the best performing data filtering metric for OCR or text-rendering related pre-training data.

4.2 MLLM Pre-Training on Mixed Image-Text Caption and Interleaved Data

Since the experimental results on caption-data pre-training in Section 4.1 have demonstrated the effectiveness and priority of UniFilter on caption data filtering, we further investigate the effectiveness of UniFilter on filtering interleaved image-text document data to promote the in-context learning capability of MLLM during multimodal pre-training. We use the OBELICS (Laurençon et al., 2024a) as the original image-text interleaved document data resource for these experiments.

Baseline and Training Setup. Since there is **no** effective data filter on filtering high-quality image-text interleaved data on document level, we consider one baseline of no filtering on interleaved data and another DFN variant baseline for processing interleaved data. For DFN variant, we follow Zhu et al. (2024) to compute the cosine similarity between each image and each text paragraph in the original interleaved document, and only discard the images of which they do not achieve 0.15 similarity threshold with any of the text paragraph within the same document. MLM-Filter baseline is deprecated here because it can only process the image-text paired caption data and cannot process the interleaved document data. As for our filtering method, the top-15% high-quality documents, as determined by the UniFilter quality scores, are selected as the training data for our MLLM. Given that each induced MLLM will be trained on a mixture of caption and interleaved data, we fix the pre-training caption data as the UniFilter curated caption data in Table 3 from DataComp-Medium for two baselines and our method to ensure there are no effects from the high-quality caption data side. And then we mix the 5B fixed caption data tokens and 5B image-text interleaved data tokens curated by three methods for each MLLM pre-training. This data mixture ratio of 1:1 is validated in MM1 (McKinzie et al., 2024) to be the optimal data mixture ratio to enhance both the multimodal few-shot and zero-shot learning. We report the 4-shot and 8-shot multimodal in-context learning performance on each VQA dataset for the baselines and our method. We select 5 random seeds for demonstration example sampling and report the mean score on 5 random seeds as the final task performance.

Results. The results of 0-shot, 4-shot and 8-shot multimodal in-context learning on 5 VQA datasets are presented in Table 4. We also provide the original results of MM1 (McKinzie et al., 2024) and BLIP-3 (Xue et al., 2024) as references even if their training size is 10-40 times larger than ours. Our method UniFilter significantly outperforms DFN variant filter baseline on GQA, VizWiz, OKVQA and TextVQA, while slightly lags behind VQA-v2 because the VQA-v2 is constructed from MSCOCO (Lin et al., 2014), which is used as the continue training data of DFN. Finally, the UniFilter induced MLLM achieves +0.7 and +2.8 average accuracy improvements over the DFN baseline on 4-shot and 8-shot in-context learning, respectively. The 0-shot in-context learning improvements are much more remarkable than that of 4-shot and 8-shot settings, achieving +3.2 average VQA task improvements. Compared with 4-shot and 8-shot in-context learning with useful instructional information and knowledge from the demonstrations, the 0-shot setting is a more challenging task which relies more on the instruction following capability of models gained from pre-training corpus. Such outstanding 0-shot improvements demonstrate the benefit of effective high-quality interleaved data filtering.

4.3 Visual Supervised Fine-Tuning

To further investigate the advantage of high-quality multimodal pre-training on the instruction-tuned MLLM, we perform visual supervised fine-tuning (SFT) on different pre-trained base MLLMs from Section 4.2. The multimodal SFT data is a joint set of visual instruction data from LLaVA-1.5 (Liu et al., 2023a) and ShareGPT4V (Chen et al., 2023a). The composition of 575k multimodal SFT data is listed in Appendix D. In addition to 5 VQA datasets, we include 4 multimodal benchmarks, POPE (Li et al., 2023b), MMMU (Val) (Yue et al., 2024), MMBench (Dev) (Liu et al., 2025), and MMStar (Chen et al., 2024) for comprehensive evaluations towards SFTed models.

Results. The evaluation results of fine-tuned MLLMs are presented in Table 5. The fine-tuned MLLM pre-trained on high-quality multimodal data curated by UniFilter significantly outperforms the instruction-tuned MLLMs with baseline filtering methods, surpassing the best baseline by +3.1 average VQA accuracy, +1.5 MMMU accuracy, and +1.6 MMBench accuracy. Further, the 0-shot VQA task performance comparisons between the

Methods	#Train Tokens	Shots	GQA	VQA-v2	VizWiz	OKVQA	TextVQA	Avg.
MM1-3B (McKinzie et al., 2024)	400B	0		46.2	15.6	26.1	29.4	-
		4	-	57.9	38.0	48.6	45.3	-
		8	-	63.6	46.4	48.4	44.6	-
MM1-7B (McKinzie et al., 2024)	400B	0		47.8	15.6	22.6	28.8	-
		4	-	60.6	37.4	46.6	44.4	-
		8	-	64.6	45.3	51.4	46.3	-
BLIP-3 (Xue et al., 2024)	100B	0	-	43.1	-	28.0	34.0	-
		4	-	66.3	-	48.9	54.2	-
		8	-	66.9	-	50.1	55.3	-
No Filtering	10B	0	17.6	22.5	12.2	23.9	29.1	21.1
		4	40.4	58.0	37.9	44.6	38.6	43.9
		8	40.7	58.6	51.5	45.5	41.0	47.4
DFN (Fang et al., 2023)	10B	0	21.8	36.6	16.7	20.6	30.4	25.2
		4	40.9	**60.0**	39.2	43.6	**43.6**	45.5
		8	41.0	**61.5**	45.9	44.9	43.9	47.4
UniFilter	10B	0	**22.9**	**37.8**	**22.4**	**25.1**	**33.9**	**28.4**
		4	**42.2**	59.7	**40.6**	**44.8**	43.5	**46.2**
		8	**42.0**	60.8	**56.3**	**46.4**	**45.5**	**50.2**

Table 4: Results of MLLMs trained on baseline data and UniFilter curated high-quality data for **10B** tokens. Each 4/8-shot accuracy value is the mean score on 5 random seeds for multimodal in-context learning evaluations.

Interleaved Pretrain Data	GQA	VQA-v2	VizWiz	OKVQA	TextVQA	VQA Avg.	POPE	MMMU Val	MMBench Dev	MMStar
No-Pretrain	28.5	49.7	15.5	27.5	32.3	30.7	81.8	40.5	74.4	36.9
DFN	32.3	57.3	16.7	27.0	41.6	35.0	82.9	39.8	75.4	38.0
UniFilter	**33.0**	**60.7**	**19.5**	**32.3**	**44.7**	**38.1**	**83.2**	**42.0**	**77.0**	**38.5**

Table 5: Zero-shot results of different instruction-tuned MLLMs on VQA datasets and multimodal benchmarks.

SFT MLLMs and their corresponding base models demonstrate all MLLMs benefit from visual SFT on completing out-of-distribution VQA tasks. The No-Pretrain (SFT-only) baseline significantly lags behind all pre-trained and fine-tuned MLLMs, demonstrating the necessity and benefit of multimodal pre-training.

5 Related Work

Data Filtering for LLM and MLLM Pre-training. The family of Phi LLMs (Abdin et al., 2024) adopt the educational value metric as the data quality metric for filtering high-quality text data for model pre-training, and FineWebEdu-Classifier (Penedo et al., 2024) is an open-source effort on training a data quality classifier for assessing the educational value of web pages. The SOTA open-sourced LLMs, Llama-3 also adopt similar data quality classifier trained on the synthetic sample-score pairs generated by Llama-2-70b, while their classifier training details are not released. In additional, DCLM (Li et al., 2024) proposes that instead of training a multi-way quality classifier, a simple binary fasttext (Joulin et al., 2016) classifier trained on positive instruction tuning data and negative web-crawled data is effective enough to curate high-quality data for SOTA LLM pre-training. In multimodal scenarios, LAION (Schuhmann et al., 2021) firstly adopts CLIPScore-based data filtering to select high-quality image-text caption data, and BLIP (Li et al., 2022) adopts the Cap-Filt data quality boosting method to generate high-quality multimodal training data.

Data Quality Classifier Trained with Synthetic Data. DCLM (Li et al., 2024) proposes to construct contrastive data for training a binary text data quality classifier by selecting LLM generated instruction data as positive data and original web-crawled data as negative data. To go beyond the synthetic binary scores, FineWebEdu-Classifier adopts Llama3-70b (Dubey et al., 2024) to generate multi-way quality scores following a well-defined human-drafted score annotation criteria. The data quality classifier to support Llama-3 pre-training also adopts a similar pipeline to instruct Llama-2-chat (Touvron et al., 2023) model to generate the quality scores. In synthetic quality score genera-

tion for multimodal data, MLM-Filter (Wang et al., 2024) prompts the GPT-4V to generate the 100-way quality scores on 4 different quality metrics to train a quality classifier for filtering image caption data from 4 distinct perspectives, while AITQE (Huang et al., 2024) simlifies MLM-Filter to one unified quality metric on a scale of 0-10.

6 Conclusion

We propose an efficient MLLM-based Unified Multimodal Data Quality Classifier to filter both high-quality image-text caption and interleaved data. Pre-training MLLMs on the high-quality data curated by the proposed UniFilter can significantly enhance the capability of these general-purpose models on downstream tasks. UniFilter overcomes the limitation of being only capable of filtering caption data in CLIP-based data filters and paves a way to steadily improve both zero-shot and few-shot multimodal in-context learning capability of pre-trained and fine-tuned MLLMs via unified multimodal high-quality data filtering.

Limitations

It is important to note that the quality of the synthetic text used for UniFilter training depends on the capabilities of the proprietary multimodal language model. Firstly, generating original 80k data takes about several hundreds of US dollars for the cost of Anthropic API. Secondly, the synthetic text in UniFilter training data will bring some potential safety concerns. We adopt the Llama-guard safety classifier model (Dubey et al., 2024) to filter out unsafe contents and ensure there is no harmful or dangerous generated texts in UniFilter training data. Future work could explore the impact of using different advanced open-source models like Qwen2.5-VL-72B (Yang et al., 2024) to generate data and achieve competing quality classification performance as the close-source MLLMs.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Anthropic. 2024. Introducing the next generation of claude. https://www.anthropic.com/news/claude-3-family. Accessed on March 4, 2024.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3).

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. 2024. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13817–13827.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. 2023. Data filtering networks. *arXiv preprint arXiv:2309.17425*.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2023. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. 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*, pages 6904–6913.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering

visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Han Huang, Yuqi Huo, Zijia Zhao, Haoyu Lu, Shu Wu, Bingning Wang, Qiang Liu, Weipeng Chen, and Liang Wang. 2024. Beyond filtering: Adaptive image-text quality enhancement for mllm pretraining. *arXiv preprint arXiv:2410.16166*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

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

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. 2024a. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024b. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, et al. 2024. Datacomp-lm: In search of the next generation of training sets for language models. *arXiv preprint arXiv:2406.11794*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

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

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pretraining for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2025. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer.

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

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. 2024. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush

Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

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

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Weizhi Wang, Khalil Mrini, Linjie Yang, Sateesh Kumar, Yu Tian, Xifeng Yan, and Heng Wang. 2024. Finetuned multimodal language models are high-quality image-text data filters. *arXiv preprint arXiv:2403.02677*.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. 2024. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. 2024. Deco: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv preprint arXiv:2405.20985*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2024. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36.

A Training Settings of MLLM Pre-Training

The training details and hyperparameters for MLLM pre-training are presented in Tab. 6. We do not follow Qwen-VL (Bai et al., 2023) to perform a separate stage to train the visual projector only. The MLLM pre-training only involves one single stage to update the parameters of visual projector and LLM backbone, while the vision encoder is frozen all the time. To accelerate the MLLM training and avoid too many padding tokens, we perform sequence packing to regroup image-text data at varied length into a fixed context size sequences. A special <|endofchunk|> token is added before the start of every image in an image-text interleaved document to indicate the end of a text paragraph. The MLLM training on 10B mixed multimodal tokens is conducted on 4 A100-40G gpus nodes, and each node contains 8 A100-40G gpus. The training for 10B tokens takes about 640 A100-40G gpu hours.

B Statistics of Multimodal Datasets

We list the statistics of the large-scale image-text caption dataset and image-text interleaved document dataset in Tab. 7 as well as their licenses. The filtered high-quality data subset by UniFilter will also inherit the original licenses of these datasets and ensure the proper usage of them. All images in two datasets are released in image-urls rather than files, leading to a large-scale invalid data samples. We discard the whole document from OBELICS if any one of the image in the document is invalid. As

Details	MLLM Pre-Training
Vision Encoder	SigLIP-so400m-384px
Visual Projector	2d Adaptive Average Pooling
LLM Backbone	Phi-3-mini-4k-instruct
Context Length	4096
Precision	BF16
Global Batch Size	256
# Training Steps	9537
# GPUs	32 A100
Peak LR	3e-5
# Warmup Steps Ratio	3%
LR Scheduler	Cosine LR Decay
Weight Decay	0.01
Adam (β_1, β_2)	(0.9, 0.98)

Table 6: Training details for MLLM pre-training on 10B multimodal tokens.

of June 2024, only a half of OBELICS interleaved documents are fully downloadable.

Dataset	#Samples	Downloadable #Samples	License
DataComp-Medium	128M	99.6M	MIT
OBELICS	141M	70.5M	CC-BY-4.0

Table 7: Pre-Training Multimodal Dataset Statistics.

C Statistics of the Curated Interleaved Document Dataset

We analyze the features and statistics of the curated interleaved document data using different filtering methods in Table 8. Compared with the No-Filter and DFN-Filter baselines, the selected high-quality document data contains more images and more text tokens in one document. Additionally, the DFN-CLIP-Filter can only remove the irrelevant images in a document based on cosine similarity, leading to an uncontrollable filtering fraction. The proposed UniFilter can achieve document-level filtering and flexibly select the filtering fraction hyperparameter based on the data quality needs.

D Visual SFT Data Composition

We select a comprehensive and diverse task sets for constructing the visual SFT instruction dataset. Since the 5 VQA datasets are used as evaluation benchmarks, the visual instruction data constructed from these 5 VQA dataset are excluded in the joint SFT dataset. For visual instructions, we select LLaVA-Conversations (Liu et al., 2023b), LLaVA-Reasoning (Liu et al., 2023b), ShareGPT4V-Caption (Chen et al., 2023a), OCRVQA (Mishra et al., 2019), A-OKVQA (Schwenk et al., 2022), TextCaps (Sidorov et al., 2020), RefCOCO (Kazemzadeh et al., 2014), and VG (Krishna et al., 2017). We use ShareGPT as the text instruction data. The final joint SFT dataset consists of 575k multimodal instructions.

Filter	Avg. #Img.	Avg. Text Len.	Avg. Document Len.	Filtering Fraction
None	1.98	842.5	1125.8	100%
DFN	1.88	841.1	1110.4	90.5%
UniFilter	3.15	1627.8	2078.3	15%

Table 8: The curated interleaved document data statistics using different filtering methods.

Data	Size	Response formatting prompts
Visual Instructions 517k		
Conversation	58K	–
Reasoning	77k	–
ShareGPT4V	100k	–
OCRVQA	80k	Answer the question using a single word or phrase.
A-OKVQA	66K	Answer with the option's letter from the given choices directly.
TextCaps	22K	Provide a one-sentence caption for the provided image.
RefCOCO	48K	*Note: randomly choose between the two formats* Provide a short description for this region.
VG	86K	Provide the bounding box coordinate of the region this sentence describes.
Text Instructions 40k		
ShareGPT	40K	–
Total 575k		

Table 9: Visual and Text SFT Data Composition.

E Full Synthetic Data Generation Prompts

The full prompts for both caption data and interleaved data generation are listed in Tab. 10. The "{multi-level quality requirements}" are placeholders for integrating the defined quality requirements in Tab. 10 into the synthetic data generation prompt.

F Ablations on In-Context Learning Prompt Templates

The demonstration prompt template affects the performance of multimodal in-context learning. We perform an ablation study on different prompt templates for constructing demonstration examples, shown in Tab. 11. The results present that the <|endofchunk|> token is significant to multimodal in-

Caption Data Generation Prompt

You are a helpful assistant to help users write two opposite image captions for the given image in JSON format. The JSON object must contain the following keys:
- "topic": a string, a topic word of this image
- "positive_caption": a string, a high-quality, comprehensive, detail-enriched caption for this image.
- "negative_caption": a string, {multi-level quality requirements}

Please adhere to the following guidelines:
- Both captions should be at least {num_words} words long.
- Both captions should be in English.
- Please avoid using complex or advanced words in the captions. Ensure that the language is suitable for a high school level audience or lower.

Your output must always be a JSON object only, do not explain yourself or output anything else. Be creative!

Interleaved Data Generation Prompt

You are an assistant to help users to write a document given several images. These images are extracted from a paper, report, or article in which these images are inserted.

<guideline> Please firstly generate a xml tag for each image in order for future generation. For each image, please generate a xml tag like "<img>image description</img>". You need to replace the image description with your generated short description of this image which is less than 5 words.

For the second task, {multi-level quality requirements}

Please adhere to the following guidelines when writing this document:
- The paragraphs in the document should be in varied length.
- The document should contain at least 500 words.
- You NEED to use xml tag as the placeholder to indicate the place where an image is inserted into.
- You NEED to ensure that all given images are used and considered.
- You MUST NOT use the image xml tag within your sentences. You should add them between sentences and paragraphs.
- You MUST use each image for ONLY ONCE in the document.

Your output must always be a JSON object only. The JSON object must contain the keys of "image_tags" and "document".

</guideline>

Now, it is your turn. Please strictly follow the above guidelines in <guideline> xml tags when writing the document.

Table 10: Prompting templates for synthetic caption data and interleaved document data generation.

context learning capability of MLLMs. The <|endofchunk|> token is inserted in the end of each text paragraph of the interleaved document data during the data pre-processing. Thus, adding this token to each demonstration example template constructs the few-shot demonstrations into an interleaved document, which may help trigger the model's parametric memory towards the pre-trained knowledge in the interleaved document data.

G Ablation Study on Filtering Fraction for Caption Data

We further investigate the effects of different fraction of retained high-quality subset from the original pool to the performance of pre-trained MLLMs. We perform ablation studies on DFN baseline and UniFilter on two filtering fractions of 15% and 30%. The results in Table 12 demonstrates that for both methods 30% filtering fraction is a better hyperparameter choice compared with retraining only 15% data. Generally, retraining less data will hurt the data diversity and distribution of curated dataset, and 30% achieves the best trade-off between highest average quality and data diversity.

H Effects of Introducing System Prompts in Multimodal In-Context Learning

In addition to the demonstration prompt template, we also investigate the effects of introducing the system prompts in the in-context learning templates. The results on these ablation studies are presented in Tab. 13. We consider 3 different system prompts as follows:

- **Phi-3 Default**: <|system|>You are a helpful assistant.<|end|>

Prompt Templates for Demonstration Examples	GQA
<|user|><image>\n{Question}Answer the question with single word or phrase.<|assistant|>{Answer}<|end|>	35.79
<bos><|user|><image>\n{Question}Answer the question with single word or phrase.<|assistant|>{Answer}<|end|>	38.46
<bos><|user|><image>\n{Question}Answer the question with single word or phrase.<|assistant|>{Answer}<|endofchunk|>	39.64
<bos><image>\n{Question}Answer the question with single word or phrase.\n{Answer}<|endofchunk|>	39.53
<bos><image>\n{Question}Answer the question with single word or phrase.\n{Answer}<|endoftext|>	34.96
<bos><image>\n{Question}Answer the question with single word or phrase.\n{Answer}<|end|>	39.09
<bos><image>\nQuestion: {Question}Answer the question with single word or phrase.\nAnswer: {Answer}<|endofchunk|>	**42.2**
<bos><image>\nQuestion: {Question}\nAnswer: {Answer}<|endofchunk|>	37.67

Table 11: Ablation studies on the effects of different in-context learning prompt construction templates on the 4-shot performance of GQA using the no-filtering baseline model.

Methods	Fraction	GQA	VQA-v2	VizWiz	OKVQA	TextVQA	Avg.
DFN	15%	25.7	35.8	21.6	24.9	30.3	27.7
DFN	30%	25.9	41.0	21.2	24.1	29.9	28.4
UniFilter	15%	26.7	41.8	20.1	24.5	28.3	29.3
UniFilter	30%	**29.6**	**43.2**	**22.9**	**28.2**	**32.5**	**31.3**

Table 12: Ablation studies on the fraction of retained high-quality subset from the original 128M data pool for MLLM pre-training.

System Prompts	GQA
Phi-3 Default	40.57
LLaVA-1.5 Default	40.77
Claude-3 Generated	41.60

Table 13: Ablation studies on the effects of different system prompt construction in templates on the 4-shot performance of GQA using the no-filtering baseline model.

- **LLaVA-1.5 Default**: A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human's questions.
- **Claude-3 Generated**: You are tasked with answering open-ended questions based on images provided from the visual question answering dataset. Each question may require understanding the visual content of the image, interpreting natural language, and applying commonsense knowledge. Your goal is to generate the most accurate answer based on the image, considering multiple possible interpretations.

Concluding from Tab. 13, introducing the system prompt in the in-context learning templates promotes the GQA performance of induced MLLM, demonstrating the success of multimodal pre-training to train the base MLLM to follow instructions.

I Examples of Claude-3 Generated Contrastive Interleaved Documents.

We provide a pair of contrastive positive and hard negative synthetic documents in Figure 4 and Figure 5. The positive document is apparently knowledge-intensive and has better detailed image-text alignment compared with the hard negative document.

In a virtual world filled with whimsical landscapes and vibrant colors, two distinctive characters stand out, their appearances strikingly contrasting yet seamlessly blending into the surroundings. One, a towering figure swathed in a snowy white ensemble with accents of vivid blue, exudes an aura of mystery and enchantment. The other, a shorter companion, dons a playful ensemble of red and blue patterns, adding a touch of vibrancy to the scene.

As they stride across the lush, emerald-hued meadows, their gazes fixate upon a solitary structure in the distance. This building, a simple yet imposing edifice, bears a bold inscription that reads 'BLOCK,' hinting at its significance within this virtual realm. The juxtaposition of the characters and the building invites speculation – are they mere explorers, or do they hold a deeper connection to this place? The surrounding environment is a tapestry of natural beauty, meticulously woven with rolling hills, scattered foliage, and a vast, azure sky that stretches endlessly overhead. Each element, from the towering trees to the delicate blades of grass, is rendered with a level of detail that blurs the line between virtual and tangible. As the characters approach the building, their purpose remains shrouded in mystery. Are they seeking knowledge, shelter, or perhaps something more enigmatic? The possibilities are boundless, and the journey promises to unveil hidden truths and unravel the secrets that lie within this captivating, digital realm.

Figure 4: A positive synthetic document generated by Claude-3. The image is sampled from OBELICS dataset.

In a distant galaxy, there exists a world where the laws of physics are vastly different from our own. Gravity works in reverse, causing objects to float away instead of falling down. The inhabitants of this bizarre realm are a race of sentient beings made entirely of stone.

These stone creatures have developed a unique culture and way of life that defies conventional logic. They construct their dwellings upside down, with the foundations on top and the roofs at the bottom. This allows them to easily float in and out of their homes, unhindered by the strange gravitational forces. Despite their rocky exteriors, the stone beings are highly intelligent and have mastered advanced technologies. They have harnessed the power of antimatter to fuel their cities, which float effortlessly in the skies above. However, their reliance on antimatter has led to a dangerous oversight – they have neglected to consider the consequences of its widespread use. Over time, the antimatter particles have begun to accumulate in the atmosphere, gradually weakening the very fabric of their reality. Cracks have started to appear in the once-seamless fabric of space-time, threatening to tear their world apart. Tragically, the stone beings remain oblivious to this impending catastrophe. They continue to go about their daily lives, blissfully unaware of the doom that looms ever closer. Their scientists, consumed by their obsession with antimatter and its applications, have failed to notice the warning signs. As the cracks in reality widen, bizarre phenomena begin to manifest. Objects and even entire structures start to phase in and out of existence, appearing and disappearing at random. The stone beings, bewildered by these occurrences, attribute them to the whims of their gods or the machinations of unseen forces. Unbeknownst to them, the very laws that govern their existence are unraveling. Soon, their world will be consumed by a void of nothingness, a fate they could have averted had they only heeded the warnings and sought alternative energy sources. In the end, their downfall will be a testament to the dangers of hubris and the importance of understanding the intricate web of cause and effect that underpins the cosmos. Their once-thriving civilization will be reduced to dust, scattered across the infinite expanse of a universe that cares not for the follies of those who fail to respect its fundamental principles.

Figure 5: A hard negative synthetic document generated by Claude-3. The image is sampled from OBELICS dataset.