AMCrawl: An Arabic Web-Scale Dataset of Interleaved Image-Text Documents and Image-Text Pairs

Shahad Aboukozzana
SDAIA, NCAI
Riyadh, Saudi Arabia
saboukozzana@ncai.gov.sa

Ahmed Ali
HUMAIN
Riyadh, Saudi Arabia
ahmed.ali@humain.ai

M Kamran J Khan
SDAIA, NCAI
Riyadh, Saudi Arabia
mkkhan@sdaia.gov.sa

Abstract

In this paper, we present the **A**rabic **M**ultimodal **Crawl** (AMCrawl), the first native-based Arabic multimodal dataset to our knowledge, derived from the Common Crawl corpus and rigorously filtered for quality and safety. Image-text pair datasets are the standard choice for pretraining multimodal large language models. However, they are often derived from image alt-text metadata, which is typically brief and context-poor, disconnecting images from their broader meaning. Although significant advances have been made in building interleaved image-text datasets for English, such as the OBELICS dataset, a substantial gap remains for native Arabic content. Our processing covered 8.6 million Arabic web pages, yielding 5.8 million associated images and 1.3 billion text tokens. The final dataset includes interleaved image-text documents and question-answer pairs, featuring 2.8 million high-quality interleaved documents and 5 million QA pairs. Alongside the dataset, we release the complete pipeline and code, ensuring reproducibility and encouraging further research and development. To demonstrate the effectiveness of AMCrawl, we introduce a publicly available native Arabic Vision Language model, trained with 13 billion parameters. These models achieve competitive results when benchmarked against publicly available datasets. AMCrawl bridges a critical gap in Arabic multimodal resources, providing a robust foundation for developing Arabic multimodal large language models and fostering advancements in this underrepresented area. **Code:** github.com/shahad-aboukozzana/AMCrawl

Figure 1: Sample QA-Image Pair from AMCrawl dataset.

1 Introduction

Multimodal Large Language Models are trained on datasets that combine multiple modalities to build models across modality understanding and generation capabilities. This led to multiple data curation efforts to build pretraining, fine-tuning, and benchmark datasets that are multimodal in nature. For Vision-Language Models, image-text pairs are among the most common and easily obtained dataset forms, since the alt-text property of images found on the web represents a quick and scalable method of finding a relevant text. However, such datasets suffer from several issues, such as an empty alt-text, alt-text filled by the image's file name, or text that is unrelated to the image content. Furthermore, if the alt-text is to be found with relevant text, the text is usually short and lacks grammatical correctness. To address this, several efforts have been made to build an interleaved image-text dataset where images appear between sequences of text. This format provides a richer and more natural context for the images; furthermore, this also exposes the model to contexts with multiple related images, which enables complex prompting scenarios involving more than one image. Multiple multimodal Large Language Models have been pretrained on interleaved multimodal documents, including Flamingo (Alayrac et al., 2022), CM3 (Aghajanyan et al., 2022), KOSMOS-1 (Huang et al., 2023) OpenFlamingo (Awadalla et al., 2023), IDEFICS (Laurençon et al., 2023), and AnyGPT (Zhan et al., 2024)

Publicly available datasets in this format are mainly targeting the English language, such as MMC4 (Zhu et al., 2023) OBELICS (Laurençon

Proceedings of The Third Arabic Natural Language Processing Conference, pages 448–465

et al., 2023) and MINT-1T(Awadalla et al., 2024). Given that and motivated by supporting multimodal LLM research for Arabic, we propose AMCrawl: An Arabic web-scale dataset of Interleaved image-text documents. The proposed dataset follows the pipeline proposed by (Laurençon et al., 2023) after customizing it for the Arabic Language. Furthermore, the pipeline is extended to generate a high quality question-answer pairs dataset, by leveraging the interleaved documents and Large Language Models. Our contributions can be summarized as follows:

- We introduce AMCrawl, a multimodal documents dataset, curated from the CommonCrawl Corpus where raw web pages are filtered for safety and quality.
- We generated a dataset of Question-Answer pairs derived from the interleaved documents using GPT generation, making it ideal to train Vision-Language Models.
- We provide a high quality Arabic translation for several multimodal datasets commonly used for training VLMs.
- We show the viability of our dataset by training and validating a Vision-Language model.
- We open source our dataset to the research community.

2 Related Works

2.1 Interleaved Image-Text Documents Datasets

Several English multimodal document datasets have been created and used to train multimodal LLMs (Zhu et al., 2023) (Raffel et al., 2020) (Laurençon et al., 2023). The Multimodal C4 (MMC4)(Zhu et al., 2023) starts from the C4 dataset (Raffel et al., 2020), downloads the images separately, then aligns image and text by solving a bipartite assignment problem for each document and its images using a CLIP model (Radford et al., 2021). OBELICS (Laurençon et al., 2023) uses recent CommonCrawl snapshots, employs the DOM structure to place images in between text sequences and de-duplicate both text and images.

MINT-1T(Awadalla et al., 2024) expanded their data sources to include PDF files and ArXiv papers. OmniCorpus (Li et al., 2024) is a multilingual interleaved image-text documents dataset covering multiple languages including Arabic, by time of this writing, the dataset is not publicly released and no statistics are provided for the Arabic portion of the dataset.

2.2 Image-Text Pairs Datasets

Peacock (Alwajih et al., 2024), a suite of Arabic multimodal large language models (MLLMs) designed to handle both vision and language tasks in Arabic. The authors also proposed Henna, a new benchmark focused on evaluating cultural and dialectal visual reasoning in Arabic contexts. Violet (Mohamed et al., 2023) is a vision–language model tailored for Arabic image captioning. The authors employed a vision encoder paired with a Gemini-based text decoder, enhancing fluency and integration between image and text representations. Image-Text pairs dataset are abundant in English, with curation processes ranging between automatic crawling of image alt-text from the web, to manual human annotation. SBU (Ordonez et al., 2011) represents one of the first efforts to collect image-text pairs at scale by querying Flickr, a social media site for image and video hosting, and filtering the results leading to 1 Million image-text pairs where the user provided description is used as a caption. MSCOCO (Microsoft Common Objects in Context)(Lin et al., 2015) is a widely used dataset offering high-quality image-text pairs, it is labeled for several tasks including object recognition, image captioning, dense pose estimation, and image segmentation. Conceptual Captions (Sharma et al., 2018) consists of large-scale image-text pairs sourced from the web, focusing on automatically generated captions with minimal human intervention. NoCaps (Agrawal et al., 2019) builds upon the COCO dataset but emphasizes evaluating models on novel object categories, encouraging generalization beyond the original dataset. LAION-400M(Schuhmann et al., 2021) and LAION-5B (Schuhmann et al., 2024) are massive datasets comprising image-text pairs scraped from the web. These datasets emphasize scalability and open-domain applications, serving as a foundation for large-scale vision-language pretraining CC12M (Changpinyo et al., 2021) is a smaller but high-quality web-crawled dataset focusing on diverse visual content and associated captions, providing a mid-scale alternative to LAION datasets Special domain datasets have been collected to address specific challenging Fashion and Lifestyle Applications DeepFashion (Liu et al., 2016) and Fashion-Gen (Rostamzadeh et al., 2018) are specialized datasets targeting fashion-related tasks, such as image captioning, clothing retrieval, and

style-based recommendations. These datasets offer detailed annotations of fashion items, including attributes, categories, and text descriptions

There are several datasets related to visual question answering and reasoning - VQA (Goyal et al., 2017), CLEVR (Johnson et al., 2017), TDIUC (Kafle and Kanan, 2017), and CVQA (Romero et al., 2024)are pivotal datasets for visual question answering. While VQA provides real-world images paired with natural language questions, CLEVR offers synthetic images designed for reasoning-based tasks. TDIUC extends this space with diverse image-question pairs, focusing on task-level diversity and difficulty.

While CVQA includes Arabic among 12 languages, it is primarily designed as a human-annotated evaluation benchmark with a focus on cultural reasoning. In contrast, AMCrawl-QA is a large-scale, automatically generated dataset containing 5 million QA pairs derived from real Arabic web documents, specifically designed for pretraining and instruction tuning of Arabic multimodal LLMs. Its integration with interleaved image-text documents enables training on long-form, multi-image contexts, making it suitable for foundational model development.

- GQA (Hudson and Manning, 2019) and VCR (Zellers et al., 2019) further explore reasoning capabilities, with GQA focusing on grounded question answering and VCR emphasizing visual commonsense reasoning in complex, multimodal scenarios.

For a fine-grained detailed understanding, several Localized Image Annotations - RefCOCO(Kazemzadeh et al., 2014) specializes in referring expression comprehension, where models must identify specific regions within an image described by natural language . - OpenImages (Localized Narratives) (Pont-Tuset et al., 2020) introduces dense annotations that include region descriptions and correspondences, aiding in tasks like visual grounding and segmentation. RedCaps (Desai et al., 2021) combines Flickr images with rich community-driven captions, offering domain-specific insights with high-quality annotations

Creative and Cultural Applications - ArtEmis (Achlioptas et al., 2021) and ArtElingo (Mohamed et al., 2022a) are datasets that focus on artistic images paired with emotional or descriptive captions, supporting research in computational aesthetics and art interpretation. - Recipe1M (Marin et al., 2019) offers text-image pairs in the culinary domain, linking recipe instructions to corresponding food images for tasks like cross-modal retrieval and generation.

To address accessibility, VizWiz (Gurari et al., 2018) provides real-world image-text pairs designed to assist visually impaired users, including visual questions and answers tailored to their needs. TextCaps (Sidorov et al., 2020) emphasizes dense text-related captioning, encouraging models to interpret and describe textual content within images, a task critical for scenarios like accessibility and information retrieval.

Multilingual datasets include: Multi30K (Elliott et al., 2016) and WIT (Wikipedia Image-Text) (Srinivasan et al., 2021) offer multilingual annotations, with Multi30K extending MSCOCO annotations to multiple languages and WIT providing image-text pairs sourced from Wikipedia across diverse domains and languages. - COYO-700M (Byeon et al., 2022) and MINT-1T (Awadalla et al., 2024) scale cross-modal datasets to hundreds of millions or billions of image-text pairs, supporting robust pretraining of vision-language models.

On the Scientific and Domain-Specific Datasets: ChartQA (Masry et al., 2022) target scientific or structured visual content, such as charts, graphs, and multimodal documents, enabling research into reasoning and interpretation in specialized domains. AI2D (Kembhavi et al., 2016) and OmniCorpus (Li et al., 2024) provide datasets for document-level image-text tasks, such as diagram understanding and multimodal document analysis. Recent large scale datasets include MMC4 (Zhu et al., 2023), OBELICS (Laurençon et al., 2023), PixelProse (Singla et al., 2024), and CommonPool (Goyal et al., 2024) are newly emerging large-scale datasets supporting diverse tasks like image captioning, dense text-image alignment, and large-scale multimodal research. There are a number of Arabic Image-Text Pairs datasets, here we review some of the most significant onces. Google's Wikipedia-based Image Text (WIT) Dataset (Srinivasan et al., 2021) is a multilingual dataset that extracts images, their captions, alt-text and attribution description, alongside a portion of the text found on the same page as a context. The Arabic subset of the dataset includes more than 600k examples covering 500K unique images. Crossmodal-3600 (Thapliyal et al., 2022) is an Image Captioning consisting of 3600 images annotated manually in 36 languages, including Arabic. ArtELingo (Mohamed et al., 2022b) is a multilingual collection of 80K artwork annotated with captions and emo-

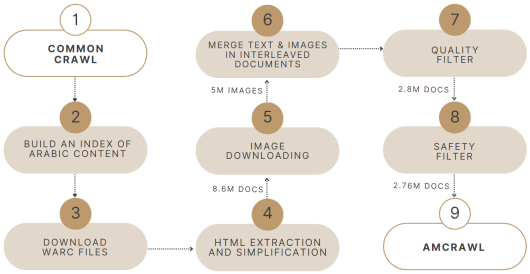


Figure 2: A high level workflow of the AMCrawl Pipeline

tions. MTVQA (Tang et al., 2024) is a multilingual Visual Question-Answer dataset covering 9 languages including Arabic and annotated by native speakers, the dataset focuses on images with textual information relevant to the questions, the dataset contains 6.68k samples, 568 of which are in Arabic. ArMeme (Alam et al., 2024) is a meme dataset consisting of 6k image-text pairs collected from social media sites. Several other efforts translate English image captioning dataset into several languages including Arabic, examples of such efforts are: Araclip (Al-Barham et al., 2024), they translate CC3M, CC12M, SBU, MSCOCO and XTD-10(Aggarwal and Kale, 2020). Table 1 summarizes key multimodal datasets used in recent research, covering dataset language, size, type, and year of release. Size refers to the number of images in image-text pairs dataset, and it refers to the number of documents in the interleaved datasets. In the language column, Multilingual+ar means that Arabic is one of the languages included in the dataset. Note that datasets are not mutually exclusive; e.g. the AraClip dataset is a translated version of several datasets like COCO.

3 Multimodal Documents Curation Process

The pipeline to process CommonCrawl data is adopted from (Laurençon et al., 2023), after several modifications and customization. The following is an overview of the main steps in the interleaved documents curation process depicted in Figure 2.

CommonCrawl Download

Using the latest CommonCrawl snapshot as the time of this writing, namely June 2024's snapshot, an index is build for metadata of a webpage with language tag equal to *ara*, for the Arabic language. Then, the metadata are used to download the Web Archive (WARC) files for the selected web pages.

Table 1: Summary of Datasets Used in Recent Multimodal Research

Dataset	Lang	Size	Type	Year
CC12M	EN	12M	Img Cap	2021
LAION	EN	400M	Img Cap	2021
RedCap	EN	12M	Img Cap	2021
ArtEmis	EN	80K	Img Cap	2021
WIT	multi	11.5M	Img Cap	2021
Crossmodal	multi	3.6K	Img Cap	2022
ArtELingo	multi	80K	Emo Pred	2022
XVNLI	multi	724K	NLI	2022
COYO-700M	EN	747M	Img Cap	2022
LAION-5B	multi	5B	Img Cap	2022
M3W (Flamingo)	EN	185M	Interleaved	2022
ChartQA	EN	20K	Chart VQA	2022
MMC4	EN	571M	Interleaved	2023
OmniCorpus	multi	8.6B	Interleaved	2023
OBELICS	EN	353M	Interleaved	2023
KOSMOS-1 Data	EN	71M	Interleaved	2023
MTVQA	multi	2K	VQA	2024
ArMeme	AR	6K	Content Filtering	2024
AraClip	AR (trans	12M	Img Cap	2024
MINT-1T	EN	3.4B	Interleaved	2024
Web Interleaved (MM1)	EN	1B	Interleaved	2024
CoMM	EN	1M	Interleaved	2024
PixelProse	EN	16M	Img Cap	2024
CommomPool	EN	12.8M	Img Cap	2024
AMCrawl (Ours)	AR	2.8M	Interleaved	2025
AMCrawl - QA (Ours)	AR	5M	VQA	2025

HTML Extraction and Simplification After the download is completed, the HTML content of the WARC file is extracted and the HTML is simplified by following several steps, including: Remove non-text or non-images nodes, Merging consecutive text nodes, Strip multiple line breaks, Strip multiple spaces, Remove HTML comments, Replace Line Break tags with line breaks, Remove dates, and simplify nested HTML nodes.

The results of this step include the simplified HTML files and URLs for all the images found in the original web page.

Image Downloading Using the image URLs from the previous step, all images are downloaded. Furthermore, a map is created between image URLs and image files.

Merging Text with Images The images are merged with the simplified text documents by replacing the image URLs with the corresponding image downloaded in the previous step. Furthermore, a basic image filtering is done at this stage where the image is not placed in the document if its URL contains one of several banned words such as *logo, button, icon, plugin, widget* to eliminate semantically irrelevant images.

Quality Filtering The previous steps produce interleaved image-text documents, which are passed through the following quality filters: removing documents with no images or more than 30 images, check image format, size and aspect ratio, check the number of words per document, and check the perplexity score for each document.

Perplexity model. We compute document-level perplexity using a **KenLM** n-gram language model trained on Wikipedia (Heafield, 2011).

Safety Filtering NSFW image filtering is done at two stage. Before downloading the images, image urls are filtered, and any url containing words related to NSFW are removed. Furthermore, downloaded images are later filtered by identifying NSFW images using an open source NSFW classifier based on the MobileNet architecture(Laborde, 2023). Any document containing at least one flagged image is removed from the dataset. After running this filter we eliminated 44,701 documents.

Table 2: General Statistics of the AMCrawl Dataset

Sr.	Category	Count
1	Downloaded Documents	8,641,036
2	Filtered Documents	2,807,179
3	Filtered Images	5,199,707
4	Documents with No Images	3.8 M
5	Train Split Images	4,496,964
5	Test Split Images	702,743
6	Total Text Tokens	1.3B

4 Data Analysis

4.1 General Statistics

As shown in Table 2 more than 8.5 million Arabic web pages were downloaded from the CommonCrawl snapshot of June 2024. Around 65% of them are eliminated in the filtration step described in Section 3. The majority of the reason for the elimination is found to be the absence of images in the document. The number of documents after the quality filter is more than 2.8 million interleaved documents.

4.2 Topic Modeling

Following (Zhu et al., 2023) we perform topic modeling using Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to understand the topic distribution and diversity across the dataset. We run LDA with 20 topics on a random subset of 1,000,000 documents, and used the learned model to infer the topics of the remaining documents. We show the frequent words and the estimated number of documents for each topic in the appendix. We observe that the documents cover a diverse set of topics including news, technology, tourism and cooking recipes. We also list the most frequent 100 domains in the appendix.

4.3 Qualitative Assessment for Dataset Samples

Following (Laurençon et al., 2023), we randomly sampled 250 documents from the interleaved dataset and manually assessed their quality and safety. Our inspection included 1,098 images, revealing that 4.2% were NSFW, 17.2% depicted logos, and 24.9% included human faces. This assessment was conducted on a sample of documents prior to the final NSFW filtering stage, as described in Section 3.5. The NSFW images were primarily associated with political topics, including visuals of protests, military operations, weapons, and ex-

plosions. These images are a natural consequence of the dataset's inclusion of articles and documents related to political news and events, which often feature such content to provide context or illustrate the subject matter. As these image-text pairs are tailored to political reporting, they reflect the inherent nature of political media and its visual representation of global events. Logo images, on the other hand, were identified as those lacking meaningful contextual relevance to the accompanying text.Importantly, any document containing NSFW content was removed in the subsequent safety filtering step; thus, the final AMCrawl does not contain NSFW material.

4.4 Dataset Viability

To test the viability of our dataset, we randomly split the dataset into training and test sets, using a 90-10 split. Each document is in both split is passed to GPT to generate question-answer pairs using a prompt following (Liu et al., 2023).

4.5 Model Architecture

There is a wide variety of multimodal LLMs in terms of architecture and functionality; however, all shared a common backbone pattern. Our model architecture in figure 3 follows the standard design of combining a visual encoder and a language model (Jin et al., 2024) (Liu et al., 2023) in the multimodal model setting. It is made up of three parts.

4.6 Image encoder:

An encoder that processes the image and generates visual tokens of the image. Vision Transformers (ViTs) are neural networks that are designed particularly for these kinds of task. Vision Transformer first splits the whole image into a sequence of fix size non-overlapping patches, then flattens those patches, and finally generates embedding vector for each flattened patch. For the image encoding task, we adopt the pre-trained CLIP (Contrastive Language-Image Pre-Training) ViT-L/14 visual encoder. (Radford et al., 2021). CLIP (Contrastive Language-Image Pre-Training) is a ViT based transformer architecture that is trained on a variety of image-text paired datasets such as MS-COCO (Lin et al., 2014). In our case CLIP image encoder processes each image individually, transforming it into the corresponding visual tokens.

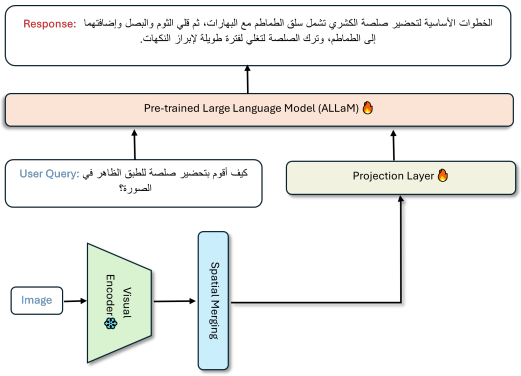


Figure 3: Overview of model architecture.

4.7 Projector:

The job of a projector is to take visual tokens from the image encoder and learn a trainable projection layer W to transform these visual tokens into language embedding tokens. There are different choices of projectors in the literature, such as MLP-based adapters (e.g., LLaVA (Liu et al., 2023)) and cross-attention projectors (e.g., Chat-UniVi (Jin et al., 2024)). We use a single linear projection layer. We opted to use a single linear projector that transforms vision tokens into the multimodal embedding space. These language tokens are fed to the large language model as input in addition to the text prompt. To avoid any mis-match the output of the projector scaled to match the input dimension of the large language model.

4.8 Large Language Model

LLMs are very large neural network architectures that are pre-trained on very large amounts of natural language data. The underlying transformer in LLM is a set of neural networks that consist of decoder blocks with self-attention capabilities. To incorporate our Arabic dataset, we used ALLaM: Arabic Large Language Model (ALLaM) (Bari et al., 2024) as LLM. The goal of ALLaM is to support the cultural values of the Arabic speaking countries. ALLaM is trained on mixed English and Arabic, in-house crawled dataset from Web documents, news articles, books (literature, religion, law and culture, among others), Wikipedia (over 1M articles), and audio transcripts (books and news). There are four different model sizes of ALLaM 7B, 13B, and 70B. We opted to use ALLaM 13B as LLM in our multi-modal setting.

A Vision-Language model based on Chat-UniVi (Jin et al., 2024) architecture is used. The model is composed of a pretrained vision encoder,

and a pretrained Language Model. Originally, Chat-UniVi (Jin et al., 2024) uses Vicuna as a Language Model. We replace it with ALLaM (Bari et al., 2024), a large language model for Arabic and English. The visual embeddings are passed to a projection layer which is trained from scratch.

5 Training Strategies

The training process of the model consists of two stages: pre-training and supervised fine-tuning training. Details of the datasets and training configuration for each stage are summarized in Table 3.

5.1 Pre-training

The pre-training phase aims to align visual and textual modalities by training the projection layer that maps visual features to the language model's embedding space. The primary goal in this stage is to optimize the projection layer while freezing the parameters of both the image encoder (CLIP ViT-L/14) and the large language model (ALLaM 13B). This ensures that the model learns to map visual tokens effectively into the language embedding space. We use large-scale, high-quality datasets such as CC3M-595K and MSCOCO. These datasets are translated into Arabic using GPT-4o to maintain linguistic and cultural consistency.

5.2 Supervised Fine-tuning

During this stage, we freeze the visual encoder and optimize the language model and adapter module. The supervised fine-tuning stage aims to enhance the model's ability to follow detailed instructions and generate accurate responses in multimodal contexts, especially within culturally specific Arabic scenarios. The fine-tuning process uses AMCrawl QA pairs, derived from our curated interleaved image-text dataset. This ensures the model is exposed to high-quality, domain-specific instructions and responses.

6 Experimental Details

The convergence of the pre-training phase (Stage 1) is illustrated in Figure 4, where the x-axis represents the number of steps and the y-axis represents the pre-training loss. It shows a steady reduction in training loss, indicative of successful alignment of the projection layer. Specifically, the loss decreased from an initial value of 7.54 to 1.74 by the end of the pretraining phase. The model learns

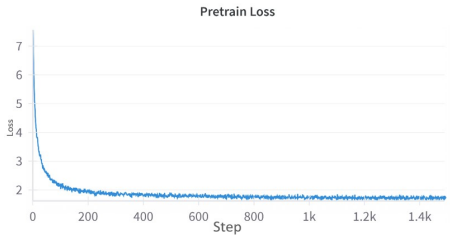


Figure 4: Pre-training Loss Convergence

to map the visual features effectively, creating a robust foundation for subsequent fine-tuning.

Figure 5 illustrates the convergence during the fine-tuning phase (Stage 2). In this phase, both the language model and the projection layer are optimized, while the vision encoder remains frozen. The loss curve demonstrates consistent improvement, indicating effective adaptation of the model to multimodal instruction tasks. The loss started at 1.77 and converged to 0.08 by the end of the second epoch.

7 Evaluation and Discussion

The evaluation data consists of 45k QA pairs randomly sampled from the test split data for cost/runtime reasons. The evaluation process is similar to LLaVA Evaluation (Liu et al., 2023), where Question-Image pairs from the test data are passed to the trained model that generate responses for each question. The responses are evaluated by asking GPT-4o to give feedback on two responses to one question: the model response and the ground truth response, GPT-4o is asked to rate the helpfulness, relevance, accuracy, level of details of the responses. Each response receives a score on a scale of 1 to 10, where a higher score indicates a better performance. The GPT-4o is also asked to provide an explanation to the generated evaluation. The evaluation results are shown on Table 4, where we evaluate the model at two stages: onces after finetuning it on translated open source multimodal data, namely MIMIC and LLaVA's 150k QA data derived from COCO, and a second time after finetuning on the training split of AMCrawl-QA. The results show a significant improvement of performance after finetuning on AMCrawl-QA.

Table 4 shows the model performance on our test set before and after finetuning using GPT scores. The results show that the model performance is enhanced by finetuning on our dataset, emphasizing the need for a dataset that reflect the culture,

Table 3: Detailed configuration for each training stage, specifying datasets, model components, and objectives.

Stage	Pre-training	Supervised Fine-tuning
Dataset	CC3M-595K, MSCOCO (Translated to Arabic)	AMCrawl QA pairs
# Samples	1.5M	5M
Trainable Components	Projection Layer Only	Projection Layer + LLM
Objective	Align visual tokens with LLM embedding space	Instruction tuning and task-specific QA

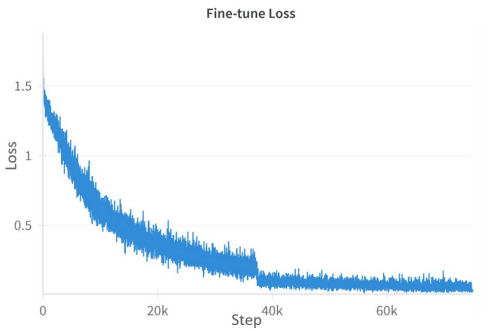


Figure 5: Fine-tuning Loss for 74K steps over AMCrawl-QA Data

Table 4: Modeling Results show performance gain after finetuning the VLM on AMCrawl. Each response was scored on a 1–10 scale by GPT-4o across helpfulness, relevance, accuracy, and detail; we report the mean.

Stage	Data	GPT Score
Finetuning-1	MIMIC+LLaVA Data	32.3
Finetuning-2	AMCrawl	43.6

traditions and history of the Arab region.

8 Ethical Considerations

The curation of AMCrawl followed strict ethical and safety measures to ensure the dataset is suitable for research use. We applied a two-stage filtering process to mitigate the risk of unsafe or harmful content. First, image URLs were screened for keywords associated with adult or explicit content, and flagged entries were excluded before downloading. Second, all downloaded images were evaluated using an open-source NSFW classifier based on the MobileNet architecture (Laborde, 2023). Any document containing at least one flagged image was removed. This procedure eliminated 44,701 documents.

Beyond safety filtering, we applied multiple quality-control steps, including removal of low-information images (e.g., logos, icons), size and aspect-ratio checks, and text perplexity thresholds. These measures ensure that the dataset prioritizes relevance, appropriateness, and linguistic quality. Importantly, the final release of AMCrawl does not contain NSFW material.

We emphasize that AMCrawl is intended strictly for academic research. While it reflects the diversity of Arabic web data, it may still inherit biases present in the original sources. We encourage users to be mindful of these limitations and to employ the dataset responsibly when training or evaluating multimodal models.

9 Limitations

While AMCrawl represents a substantial step toward building native Arabic multimodal resources, several limitations remain. The current release is derived from a single CommonCrawl snapshot (June 2024), which may not fully capture temporal or regional diversity in Arabic web content. Despite extensive filtering, residual boilerplate, redundant passages, and near-duplicate images may persist, with a preliminary perceptual hashing analysis suggesting only 45% image uniqueness. Evaluation relied on GPT-4o as an automatic judge over a sampled subset of the test data, which, while practical, differs from human annotation and may not align with results obtained using alternative evaluators or standardized benchmarks; ongoing work includes testing on Henna, CVQA, and other Arabic VLM benchmarks. Comparisons to existing Arabic VLMs such as Peacock and Violet are therefore indicative rather than conclusive, given differences in dataset scale, annotation style, and evaluation protocols. In addition, large-scale translated datasets (e.g., CC3M, MSCOCO) used in pre-training may still contain translation artifacts or cultural mismatches despite GPT-4o-based translation and filtering. Finally, as AMCrawl is drawn from publicly available Arabic web data, it may inherit societal biases, uneven regional representation, or content gaps. While explicit NSFW material was removed through a two-stage filtering pipeline, other forms of bias (political, cultural, or

gendered) remain possible. These factors frame AMCrawl as a strong first release that we intend to expand and refine in future work.

10 Future work & Conclusions

We introduce AMCrawl, a multimodal dataset consisting of filtered interleaved image-text documents and image-text Question-Answer Pairs derived from the CommonCrawl. We show that such a dataset is necessary to train Vision-Language Models, and depending solely on translating image-text pairs leads to low performance on questions that require knowledge of Arabic culture and traditions. Opening such a dataset to the public enriches the multimodal Arabic dataset landscape and ensures that Arabic is well-supported in the development of Multimodal Large Language Models (LLMs).

The findings of this work show the viability of collecting large-scale multimodal web data for training Multimodal LLMs. While this study was run on a single CommonCrawl Snapshot, which represents one month's worth of web crawl data, future work aims to scale the pipeline to cover a wider time window and generate a higher volume of data. Another promising direction for future work is to train a native-Arabic CLIP model. This serves two purposes: on one hand, an Arabic CLIP model serves as an evaluator and a quality filter for Image-Text association and relatedness, leading to more semantically related image-text pairs. On the other hand, the current dominant approach in building Vision-Language Models is to use a pretrained Image Encoder alongside a pretrained LLM (Laurençon et al., 2024). A native-Arabic CLIP backbone can provide better visual embeddings for Arabic multimodal models.

In addition, future work will address several limitations identified in this study by expanding beyond a single CommonCrawl snapshot, conducting domain-level geographic analysis, and incorporating deduplication and machine-generated content detection to improve data quality. We also plan to complement GPT-based judging with human evaluation, and benchmark against existing Arabic multimodal resources such as Peacock, Violet, Henna, and CVQA for a more comprehensive comparison. Finally, we will refine the large-scale translation pipeline for datasets such as CC3M and MSCOCO with additional validation to reduce residual noise.

References

Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. 2021. Artemis: Affective language for visual art. *CoRR*, abs/2101.07396.

Pranav Aggarwal and Ajinkya Kale. 2020. Towards zero-shot cross-lingual image retrieval. *Preprint*, arXiv:2012.05107.

Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, and Luke Zettlemoyer. 2022. Cm3: A causal masked multimodal model of the internet. *Preprint*, arXiv:2201.07520.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957.

Muhammad Al-Barham, Imad Afyouni, Khalid Almubarak, Ashraf Elnagar, Ayad Turky, and Ibrahim Hashem. 2024. AraCLIP: Cross-lingual learning for effective Arabic image retrieval. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 102–110, Bangkok, Thailand. Association for Computational Linguistics.

Firoj Alam, Abul Hasnat, Fatema Ahmed, Md Arid Hasan, and Maram Hasanain. 2024. ArMeme: Propagandistic content in arabic memes. *arXiv: 2406.03916*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. *Preprint*, arXiv:2204.14198.

Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. 2024. Peacock: A family of Arabic multimodal large language models and benchmarks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12753–12776, Bangkok, Thailand. Association for Computational Linguistics.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *Preprint*, arXiv:2308.01390.

Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Etash Kumar Guha, Matt Jordan, Sheng Shen, Mohamed Awadalla, Silvio Savarese, Caiming Xiong, Ran Xu, Yejin Choi, and Ludwig Schmidt. 2024. Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens. *Preprint*, arXiv:2406.11271.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan AlRashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, and 6 others. 2024. Allam: Large language models for arabic and english. *Preprint*, arXiv:2407.15390.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *Preprint*, arXiv:2102.08981.

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. 2021. RedCaps: Web-curated image-text data created by the people, for the people. In *NeurIPS Datasets and Benchmarks*.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *Preprint*, arXiv:1605.00459.

Sachin Goyal, Pratyush Maini, Zachary Chase Lipton, Aditi Raghunathan, and J Zico Kolter. 2024. The science of data filtering: Data curation cannot be compute agnostic. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

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 *Conference on Computer Vision and Pattern Recognition (CVPR)*.

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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3608–3617, Los Alamitos, CA, USA. IEEE Computer Society.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language is not all you need: Aligning perception with language models. *Preprint*, arXiv:2302.14045.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *Preprint*, arXiv:2311.08046.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997.

Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1983–1991.

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, Doha, Qatar. Association for Computational Linguistics.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision – ECCV 2016*, pages 235–251, Cham. Springer International Publishing.

Gant Laborde. 2023. Deep neural network for nsfw detection. GitHub. Accessed: 2025-04-10.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. *Preprint*, arXiv:2408.12637.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Preprint*, arXiv:2306.16527.

Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, Guanzhou Chen, Yinan He, Zhangwei Gao, Erfei Cui, Jiashuo Yu, Hao Tian, Jiasheng Zhou, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, and 21 others. 2024. Omnicorpus: A unified multimodal corpus of 10 billion-level images interleaved with text. *Preprint*, arXiv:2406.08418.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

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*, pages 740–755, Cham. Springer International Publishing.

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

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2019. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Trans. Pattern Anal. Mach. Intell.*

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Abdelrahman Mohamed, Fakhraddin Alwajih, El Moatez Billah Nagoudi, Alcides Alcoba Inciarte, and Muhammad Abdul-Mageed. 2023. Violet: A vision-language model for arabic image captioning with gemini decoder. *arXiv preprint arXiv:2311.08844*.

Youssef Mohamed, Mohamed Abdelfattah, Shyma Alhuwaider, Feifan Li, Xiangliang Zhang, Kenneth Church, and Mohamed Elhoseiny. 2022a. ArtELingo: A million emotion annotations of WikiArt with emphasis on diversity over language and culture. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8770–8785, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Youssef Mohamed, Mohamed Abdelfattah, Shyma Alhuwaider, Feifan Li, Xiangliang Zhang, Kenneth Ward Church, and Mohamed Elhoseiny. 2022b. Artelingo: A million emotion annotations of wikiart with emphasis on diversity over language and culture. *Preprint*, arXiv:2211.10780.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *ECCV*.

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. *Preprint*, arXiv:2103.00020.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 56 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *Preprint*, arXiv:2406.05967.

Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. 2018. Fashion-gen: The generative fashion dataset and challenge. *Preprint*, arXiv:1806.08317.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2024. Laion-5b: an open large-scale dataset for training next generation image-text models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

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*, abs/2111.02114.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *European Conference on Computer Vision (ECCV)*.

Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. 2024. From pixels to prose: A large dataset of dense image captions. *ArXiv*, abs/2406.10328.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. *arXiv preprint arXiv:2103.01913*.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, Yanjie Wang, Yuliang Liu, Hao Liu, Xiang Bai, and Can Huang. 2024. Mtvqa: Benchmarking multilingual text-centric visual question answering. *Preprint*, arXiv:2405.11985.

Ashish V. Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. *Preprint*, arXiv:2205.12522.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yugang Jiang, and Xipeng Qiu. 2024. Anygpt: Unified multimodal llm with discrete sequence modeling. *Preprint*, arXiv:2402.12226.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Preprint*, arXiv:2304.06939.

A Appendix

الحياة البحرية في كاوست

تُطل كاوست على شاطئ البحر الأحمر الذي يحيط به نظام بيئي بحري طبيعي وجميل. ويدرك مجتمع الجامعة بأكمله أهمية هذا النظام البيئي الذي يعتبره جزءاً لا يتجزأ من رسالتنا مسؤوليتنا التعليمية والاجتماعية. وهذا النظام البيئي منصوص في سياسة الإشراف البيئي لكاوست التي تؤكد على أن "حماية البيئة البحرية الثمينة التي تحيط بالجامعة" هو من الأهداف الرئيسية لكاوست.

تزخر المياه الساحلية المحيطة بكاوست بالشعاب المرجانية وغابات المنغروف ومروج الأعشاب البحرية والطحالب الضخمة. وتُعدّ هذه الموائل الطبيعية مختبراً حياً يستعين به العلماء في تطوير طرق جديدة للحفاظ على البيئات البحرية والساحلية.

حددت كاوست، عبر مسح أساسي للبيئة البحرية ركز بوجه خاص على المناطق الساحلية المجاورة للجامعة، جملة من الموائل الطبيعية ضمت رمال المد والجزر وسهول طينية وغابات المنغروف والأعشاب بحرية والسبخات (مسطحات مالحة فوق مستوى المياه الجوفية تماماً) وشواطئ رملية وصخرية وطحالب ضخمة ومجمعات مرجانية في مناطق المدّ والجزر والجزء العلوي من منطقة ما تحت المد.

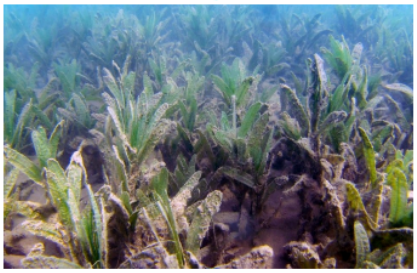
يطلق عادة على المرجان "الغابات المطيرة للبحر" لأن الشعاب التي تشكلها تُعدّ من أكثر الأنظمة البيئية تنوعاً حيوياً في العالم، وملاذاً ملائماً يوفر الطعام والمأوى لكائنات حية كثيرة تعتمد في بقائها على هذه الأنظمة البيئية. وتوفر الشعاب المرجانية أيضاً فرصاً هائلة للتعليم والاستجمام.

حددت دراسة عن المناطق البحرية المحيطة بكاوست 181 نوعاً من المرجان. وتنتمي الأنواع المألوفة إلى أجناس قميات المسام والديبساستريا (الفافيا سابقاً) والمونتيبورا والغونيبورا والفافيتيس.

وسجلت الدراسة أيضاً شعاباً مرجانية متنوعة سليمة بعيدة قليلاً عن كاوست في المياه المفتوحة للبحر الأحمر.

تتنوع الأسماك في شعابنا تنوعاً مذهلاً. وتتراوح أنواعها بين الأسماك المفترسة الضخمة كأسماك القرش والهامور والباراكودا التي تفترس أنواعاً أخرى، وأسماك صغيرة كسمك المهرج والسمك الملائكي تتغذى على العوالق والطحالب والمواد الغذائية الصغيرة. وتجذب هذه الأسماك الصغيرة زاهية الألوان الناس أيضاً إلى الشعاب بهدف التعليم والاستجمام.

سُجّل نحو 136 نوعاً من الأسماك في مواقع الشعاب المحيطة بكاوست. يشار إلى أن الكيدميات (أسماك الرأس) هي الأنواع الغالبة تليها سمكة الدامسل (السمكة الآنسة) وأسماك الببغاء (الحريد) وفراشات البحر والهامور والجراح. كما رُصدت مجموعات من الدلافين في المياه المحيطة بنا.



تُعدّ مروج الأعشاب البحرية علامة فارقة للأنظمة البيئية في السواحل الضحلة. ففضلاً عن توفير الأغذية الأساسية والمأوى للحياة البحرية، تعمل جذور هذه النباتات المزهرة على تثبيت الرواسب في مكانها، وتحد من تآكل الخط الساحلي.

سُجّل نوعان مميزان من مروج الأعشاب البحرية على امتداد شاطئ كاوست، أحدهما بالقرب من منارة كاوست، والآخر بالقرب من معلم الملك عبدالله التذكاري وهما محب الملح البيضوي (نجيل بحري) والهالودول ضيق الأوراق.

ينمو المنغروف، وهو نبات يتحمل الملوحة ذو جذور خشبية، على امتداد المياه الساحلية الضحلة. وتوفر هذه النباتات البرمائية، التي تضع قدماً في البر وقدماً أخرى في المياه، الغذاء والمأوى والموطن الحاضن لحيوانات كثيرة، ومنها الطيور وسرطان البحر والسحالي والروبيان والرخويات (الحلزون) والأسماك.

أنواع المنغروف السائدة في كاوست هي المنغروف البحري، الذي يعرف أيضاً بالمنغروف الرمادي أو الأبيض، لأن بلورات الملح تغطي أوراقه وسيقانه. أما المانغروف الأحمر، فهو نوع منفصل يوجد في منطقة صغيرة بالقرب من الشاطئ الجنوبي للجامعة. تعرف أكثر على منغروف كاوست هنا.

طبيعتنا، على الهواء مباشرة

يعرض البث المباشر لكاميرا كاوست "فيش كام" (FISH KAM) الحياة البحرية في البحر الأحمر في الزمن الحقيقي أثناء ساعات النهار، ويمنح مجتمعنا (وأي شخص آخر في العالم) نافذة على التنوع الحيوي المصان المذهل في مياه كاوست المحمية. لقد تابع المشاهدون في جميع أنحاء العالم أسماك الشعاب وأسماك اللقيطة والسلاحف وحيوانات الأخطبوط عبر هذه العدسة.

Figure 6: Example of a multimodal document (Appendix).

طريقة عمل عجينة القطايف وطريقة عمل القطايف

1- إخلطى السميد والنشاء والدقيق وماء الزهر والسكر والبيكينج بودر والماء جيداً حتى تكون لديكِ عجينة لينة وناعمة، كما يمكنك خلط المقادير السابقة فى الخلاط الكهربائى لتسهيل المهمة. وأيضا يمكنك إضافة ملعقة كبيرة من لبن البودرة على المزيج لتحلي طعمها اكثر.
2- إتركى العجينة حتى تتخمر لمدة لا تقل عن ساعة. ثم ضعيها فى إبريق صغير لتسهيل سكب المزيج منها.
3- أحضري صينية نظيفة وعليها منشفة جافة لوضع أقراص القطايف عليها بعد تمام النضج.
4- سخني مقلاة سميكة القاعدة جيداً على نار متوسطة، وإبدأي بعدها بتحضير مقدار ملعقة كبيرة من المزيج وإسكبيه فى المقلاة على شكل قرص صغير.
5- عندما تتكون ثقوب أو فقاقيع على سطح الأقراص فهذا دليل على عملية نجاح التخمر، وعندما يبدأ يجف سطحها ويحمر لونها من جهة واحدة، وقتها عليكِ بإزالتها من المقلاة ووضعها على الصينية التى قد أحضرتيها.
6- أكملي سكب بقية العجينة فى المقلاة بنفس الطريقة فى الخطوة 4.
7- والآن دعي القطايف تبرد تماماً قبل أن تبدأي فى حشوها، ولا تنسي أن القطائف يجب أن يتم حشوها وهي طازجة أو في نفس اليوم الذى صنعت فيه، لأن عند تخزينها لن تستطيعي إغلاقها على الحشوة.
8- بعد حشو القطائف يمكنك قليها فى الزيت الخفيف والتمتع بطعمها الرائع بعد وضعها فى مزيج العسل.

طريقة أخرى لطريقة عمل عجينة القطايف

- ضعي الدقيق والسميد والسكر والنشاء و"البايكنغ باودر" والماء في ابريق الخلاّط، وشغّلي الخلاط على سرعةٍ متوسّطةٍ، لتتكوّن لديكِ عجينةٌ ناعمةٌ وسائلة، اتركي العجينة ترتاح لحوالي 15 دقيقة.
- أحضري صاجاً سميكاً، أو مقلاةً سميكة القاعدة، ضعي الصاج على نارٍ متوسّطة، ليسخن.
- ضعي عجينة القطايف في كوبٍ أو في ابريقٍ صغير، لتسهيل سكبها.أحضري صينيّةً قصيرة الحافّة، وضعي عليها فوطةً قطنيّةً نظيفةً، واتركيها جانباً.
- أسكبي العجينة على شكل أقراص صغيرة، انتظري لتتكوّن ثقوب على سطح الأقراص ويجفّ سطحها. استعملي ملعقة عريضة، أنقلي الأقراص بخفّة على الفوطة ودعيها لتبرد. أكملي سكب بقيّة العجينة، ودعي القطايف تبرد تماماً قبل التشكيل.

والآن بعد أن تعلمنا طريقة عمل عجينة القطايف يجب أن نتعلم كيف نقوم بعمل القطايف !

1. العجينة: في إبريق الخلاط ضعي الدقيق، السميد، السكر، الخميرة، البيكنج باودر، الملح، الماء وماء الزهر، شغلي على سرعة متوسطة إلى أن تحصلي على عجينة سائلة القوام.
2. دعي العجينة في مكان دافئ إلى أن ترتاح لمدة ٣ ساعات مع التقليب بين الحين والآخر لتفرغي العجينة من الفقاعات المتكونة فيها أثناء التخمير.
3. سخني صاج سميك على نار متوسطة ليصبح ساخنا جدا. أحضري صينية واسعة وضعي فيها فوطة قطنية نظيفة.
4. أسكبي عجينة القطايف على الصاج لتكوني قرص متوسط الحجم أو حسب المقاس الذي تفضلين. أتركي القرص على الصاج بدون تحريك أو تقليب إلى أن تتكون فقاعات على سطح القرص وأتركيه إلى أن يجف السطح تماما. باستعمال ملعقة معدنية عريضة أنقلي قرص القطايف على الفوطة في الصينية. أكملي سكب بقية الأقراص لتنتهي كمية العجينة. غطي أقراص القطايف بالفوطة لحين تحضير الحشو.
5. القشطة: في قدر سميك القاعدة متوسط الحجم ضعي الحليب، الكريمة، الدقيق، النشا والسكر، قلبي المواد بمضرب شبك يدوي ليذوب النشا والسكر، ضعي القدر على نار متوسطة إلى أن تغلي القشطة وتصبح سميكة القوام. دعي القشطة تطهى لمدة دقيقة أو دقيقتين إلى أن تسمك وتتجانس. دعي القشطة تبرد تماما قبل الإستعمال لحشو القطايف.
6. المكسرات: في طبق عميق ضعي البندق، الجوز، الزبيب والقرفة، قلبي المواد لتختلط.
7. أمسكي قرص من القطايف وضعي في وسطه مقدار ملعقة كبيرة من القشطة أو المكسرات، أقفلي قرص القطايف على الحشو لتحصلي على شكل نصف دائرة. أكملي حشو بقية الأقراص بالقشطة والمكسرات.
8. للقلي: في مقلاة عميقة ضعي السمن والزيت بحيث يكون بارتفاع ٢ بوصة تقريبا. ضعي المقلاة على نار متوسطة ليسخن الخليط.
9. ضعي عدة أقراص من القطايف المحشوة في الزيت الساخن وإقليها لتصبح ذهبية اللون.
10. أخرجي القطايف من الزيت وضعي مباشرة في القطر. إنتظري عدة دقايق ثم أخرجي أقراص القطايف من القطر وقدميها ساخنة.

Figure 7: Another example of a multimodal document (Appendix).

Table 5: Results of LDA with 20 Topics (1M documents).

No.	Topic Label	Ratio	Related Words
2	Festivals	2.08%	السعودية، نيوز، العربية، المملكة، مهرجان، السعودي، العالم، حفل، الرياض، العالمي، جديدة، المزيد، المهرجان، السينما، الأمير، الأول، العربي، الفن، فيلم، الشعر
3	Politics	8.07%	رئيس، العراق، مجلس، اليمن، الرئيس، الحكومة، الشعب، السياسية، العام، ولد، لبنان، الوطني، الدولة، وزير، حزب، المجلس، الجمهورية، السياسي، الانتخابات، الوطنية
4	Services	3.20%	خصم، العمل، موقع، كود، وزارة، الخدمات، الاجتماعية، تقديم، الخاصة، الطبية، الأسنان، الصحية، الاجتماعي، المملكة، رقم، خدمة، طلب
5	Education	6.15%	وظائف، التعليم، جامعة، التربية، اللغة، الصف، الثالث، الجامعة، العامة، الطلاب، للصف، الفصل، الثانوية، العربية، الأول، رقم، كلية، الدراسي، التعليمية، وزارة
6	Books	4.13%	كتاب، كتب، تفسير، تاريخ، العربية، المنام، العربي، الكتب، طرف، رؤية، حلم، الشيخ، المنتدى، رواية، تحميل، برج، الحبيب، الكاتب
7	Diplomacy	9.54%	مصر، رئيس، وزير، المصرية، مجلس، العربية، الإمارات، التعاون، الدكتور، الدولي، الرئيس، العامة، المصري، وزارة، العمل، دبي، التنمية، الدولة
8	Conflict	9.87%	غزة، الاحتلال، إسرائيل، المتحدة، الحرب، الإسرائيلي، سوريا، الجيش، الفلسطينية، قطاع، حماس، قوات، فلسطين، الرئيس، مدينة، روسيا، الفلسطيني، الأمن، إيران، لبنان
9	Religion	5.92%	يا، الناس، السلام، القرآن، الحياة، شيء، يقول، الكريم، تعالى، صلى، الأرض، العالم، كنت، الإمام، وسلم، قصة، الإنسان، لقد، سورة
10	Law	2.43%	القانون، الإسلامية، قانون، عبد، الإسلام، الحج، الإسلامي، رمضان، القانونية، الإنسان، العامة، الدين، المسلمين، المحكمة، حكم، حقوق، رقم، الشيخ
11	Cleaning	3.58%	شركة، تنظيف، بالرياض، الرياض، افضل، تركيب، الكويت، نقل، صيانة، المياه، خدمة، خدمات، أفضل، شركات، الشركة، فني، عزل، مكافحة، مظلات، جدة
12	Health	4.69%	الجسم، الدم، يجب، علاج، تناول، الصحة، الأطفال، العلاج، عملية، الصحية، القلب، الحمل، صحة، فوائد، الطفل، حالة، الإصابة، العديد
13	Application	4.99%	تحميل، تطبيق، برنامج، يمكنك، التطبيق، لعبة، الموقع، مواقع، بك، مجانا، الهاتف، حساب، الدخول، تسجيل، الخاص، قم، تصميم، تطبيقات، الفيديو
14	Marketing	5.18%	يمكنك، أفضل، إضافة، استخدام، الشعر، المنتج، التسويق، يتم، الخاصة، كيفية، عرض، مجموعة، الإنترنت، العمل، العديد، تصميم
15	Economy	4.80%	أسعار، سعر، العام، ارتفاع، النفط، المالية، الدولار، بنسبة، الذهب، البنك، العالم، المركزي، شركة، الاقتصاد، السوق، الصين، زيادة، الحكومة، السعودية، المتحدة
16	Cooking	2.19%	طريقة، عمل، زيت، صور، مطعم، دكتور، عيد، العنوان، الطعام، تحضير، رمضان، الزيتون، كيلو، كوب، كبيرة، ملعقة، القهوة، أفضل، وصفات
17	Cars	4.92%	شركة، الشركة، السيارات، الاصطناعي، الرقمية، السيارة، سيارة، الذكاء، الجديدة، نظام، الشركات، أفضل، الطاقة، يتم، مجموعة، سيارات، البيانات، العملاء، الكهربائية، العملات
18	Tourism	4.93%	المغرب، عروض، مدينة، السياحة، المدينة، المغربية، مركز، الوطني، صيانة، الجزائر، السياحية، القاهرة، العالم، الجديدة، منطقة، المغربي، البحر، الوطنية، عيد
19	Sports	6.90%	مباراة، الدوري، كأس، الأهلي، القدم، الزمالك، العالم، نادي، دوري، منتخب، المباراة، الاتحاد، فريق، الفريق، كرة، المنتخب، لكرة، مدريد، مباريات، بطولة
20	Entertainment	4.49%	الحلقة، مسلسل، مصر، عبد، أحمد، فيديو، وفاة، فيلم، الفنان، رمضان، شاهد، أخبار، حلقة، عيد، مشاهدة، المصري، تفاصيل، محمود، المصرية، عرض

1.1 Most Frequent Domains

Table 6: Ranking the 100 most frequent domains in terms of number of documents (split into two sets of 50).

Rank	Domain Name	Docs	Rank	Domain Name	Docs
1	royanews.tv	23,597	51	raseef22.net	3,105
2	nn.najah.edu	12,087	52	www.masrawy.com	3,031
3	aawsat.com	9,870	53	www.alaraby.co.uk	3,014
4	hayah.cc	9,409	54	www.independentarabia.com	3,001
5	www.mxawi.com	9,125	55	altaj.news	2,981
6	www.filgoal.com	8,421	56	live.shrgiah.net	2,881
7	alwahdanews.ae	7,829	57	www.dampress.net	2,848
8	www.hayah.cc	7,783	58	www.aletihad.ae	2,843
9	observeriraq.net	7,158	59	www.alroeya.com	2,840
10	ar.hibapress.com	7,035	60	www.alrasheedmedia.com	2,809
11	www.syria.tv	6,600	61	www.elkhabar.com	2,805
12	sanews.pythonanywhere.com	6,498	62	www.sayidaty.net	2,773
13	www.akhbaralaan.net	6,484	63	www.copanetarab.com	2,746
14	thenationpress.net	6,221	64	www.yallakora.com	2,714
15	alghad.com	6,026	65	rassd.com	2,686
16	ar.lesiteinfo.com	5,747	66	smc.gov.ye	2,668
17	ekshef.com	5,487	67	www.kurdistan24.net	2,649
18	islamonline.net	5,436	68	felesteen.news	2,582
19	www.amsebehm2017.com	5,418	69	wasfetmama.com	2,569
20	www.bezaat.com	4,826	70	elmeezan.com	2,553
21	sca.sa	4,613	71	www.tabnak.ir	2,540
22	imamhussain.org	4,590	72	almesryoon.com	2,528
23	al-ain.com	4,578	73	osnplus.com	2,521
24	www.youm7.com	4,377	74	elbashayer.com	2,497
25	saharamedias.net	4,133	75	yemen-anbaa.com	2,486
26	ralia.lesiteinfo.com	4,121	76	sawaleif.com	2,485
27	www.alamatonj.com	4,060	77	www.elbotola.com	2,482
28	news.radioalgerie.dz	3,950	78	26sep.net	2,474
29	arabic.rt.com	3,928	79	islamarchive.cc	2,461
30	mwadah.com	3,906	80	ahram-canada.com	2,448
31	www.enabbaladi.net	3,786	81	www.sada-elarab.com	2,413
32	www.elwatannews.com	3,770	82	shabiba.com	2,377
33	www.alwatan.com.sa	3,753	83	www.alsirah.com	2,370
34	www.maannews.net	3,642	84	www.copts-united.com	2,353
35	www.abjjad.com	3,592	85	lakome2.com	2,353
36	thenewkhalij.news	3,550	86	slaati.com	2,353
37	www.almadenahnews.com	3,474	87	sa.aqar.fm	2,347
38	www.mobtada.com	3,447	88	hiragate.com	2,332
39	almessa.gomhuriaonline.com	3,445	89	www.dalil1808080.com	2,299
40	www.lisanerab.com	3,387	90	doctors.dailymedicalinfo.com	2,232
41	www.elaosboa.com	3,379	91	www.elzmannews.com	2,193

Continued on next page

Table 6 – continued from previous page

Rank	Domain Name	Docs	Rank	Domain Name	Docs
42	www.i7lm.com	3,345	92	www.sabr.cc	2,177
43	www.raya.com	3,319	93	www.jordanzad.com	2,169
44	digital-aarena.com	3,213	94	www.elmwatin.com	2,163
45	bezaat.com	3,185	95	www.365scores.com	2,157
46	www.soutalomma.com	3,178	96	www.albiladpress.com	2,141
47	trend.m7et.com	3,174	97	catalog.arzagplus.com	2,141
48	www.wataninet.com	3,156	98	koorasudan.net	2,113
49	www.shfanews.net	3,156	99	www.royanews.tv	2,111
50	www.skynewsarabia.com	3,149	100	nahrainnet.net	2,110

1.2 GPT Prompt for QA Generation based on a Web Document

You are an AI visual assistant, and you are seeing a single image. You receive an article describing the same image you are observing.
Answer all questions as you are seeing the image.
Design a conversation between you and a person asking about this photo. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question.
Ask diverse questions and give corresponding answers.
Include questions asking about the visual content of the image. Only include questions that have definite answers:
(1) one can see the content in the image that the question asks about and can answer confidently;
(2) one can determine confidently from the image that it is not in the image.
Do not ask any question that cannot be answered confidently.
Also include complex questions that are relevant to the content in the image, for example, asking about background knowledge of the objects in the image, asking to discuss about events happening in the image, etc. Again, do not ask about uncertain details.
Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary.

1.3 Example QA on Test Set Images

This section presents two visual QA examples from the test set. The model answers based solely on the visual content.

Figure 8: Sample 1: Test Image

Question

ما هو اسم المعلم السياحي الذي يظهر في الصورة؟

Model Answer

المعلم السياحي الذي يظهر في الصورة هو مدينة تبليسي، العاصمة الجورجية.

Figure 9: Sample 2: Test Image

Question

ما هو اسم المسجد الذي يظهر في الصورة ومن قام ببنائه؟

Model Answer

اسم المسجد هو مسجد المشور، وتم بناؤه على يد أبي حمو موسى الأول.