# A Simple Baseline for Zero-shot Visual Question Answering via Synthetic Data Generation

Anonymous ACL submission

## Abstract

Zero-shot Visual Question Answering (VQA) poses a challenging and crucial task in vision and language reasoning, demanding models to generate answers based on questions and images without human annotation. Previous approaches mainly focus on transforming images into captions and utilizing language model knowledge to answer visual questions. Despite the promising results, such a paradigm suffers from hallucination and high inference costs. In this paper, we propose a zero-shot VQA framework MKDG, which transfers knowledge from large language models (LLMs) and multi-modality models through a synthetic data generation approach, thus utilizing the ability of LLMs and mitigating the hallucination. Specifically, our method introduces a three-step synthetic data generation and training pipeline that first creates pseudo questions and answers with caption model and LLMs. To alleviate the hallucination and unbalanced data distribution in synthetic data, we propose a CLIP-based filtering and data selection strategy. Finally, we fine-tune a moderate-sized generative vision language model with the automatically curated synthetic dataset to perform VQA task. Experimental results on popular VQA benchmarks demonstrate the effectiveness of MKDG. We achieve superior performance and outperform outperforming strong baselines incorporating GPT-3 with significantly lower inference cost.

## 1 Introduction

Visual Question Answering (VQA) is a core challenge in vision and language (VL) tasks, which aims to understand and answer questions related to visual inputs. It plays a crucial role in complex real-world applications such as visual dialog (Das et al., 2017), visual relationship detection (Lu et al., 2016) and vision language navigation (Anderson et al., 2018). However, training a robust and versatile VQA model traditionally demands a substantial amount of human annotations, which can be costly and introduce various data biases. To address this, a promising strategy is to achieve zero-shot learning for a specific VQA task by exploiting rich language/image data from related VL tasks (Lin et al., 2014), or transferring knowledge from pre-trained large language models (LLMs) (Zhang et al., 2022; Brown et al., 2020a; Touvron et al., 2023a,b). Those generic knowledge sources make it possible to accomplish VQA tasks without human-annotated question-answers.

Previous explorations in zero-shot VQA can be summarized into two categories. The first paradigm leverages pre-trained foundation models to directly perform zero-shot VQA. Recent studies (Yang et al., 2022; Hu et al., 2022; Tiong et al., 2022; Guo et al., 2023; Du et al., 2023) use a two-stage approach with Large Language Models (LLMs): representing visual inputs through captions and then generating answers with an LLM. This method benefits from the reasoning capability of LLMs but faces *visual information loss* due to the inherent limitations of captioning and *expensive inference* with large LLMs like GPT-3 (Brown et al., 2020a). An alternative approach (Li et al., 2023; Awadalla et al., 2023; Alayrac et al., 2022) aligns LLMs (Zhang et al., 2022; Chung et al., 2022a) with visual encoders (Dosovitskiy et al., 2020; He et al., 2015) using image-caption data. While this method demonstrates promising generalization in VQA tasks, it is prone to *restricted reasoning capability*. The alignments in these models focus on describing the visual scene rather than extracting visual information relevant to the question. The second paradigm generates synthetic VQA samples (Banerjee et al., 2020; Changpinyo et al., 2022) by converting existing image-caption pairs into image-question-answer triplets. However, these approaches often suffer from *deficiency in information* due to limited caption content.

To address these challenges, we propose a novel data generation framework for zero-shot

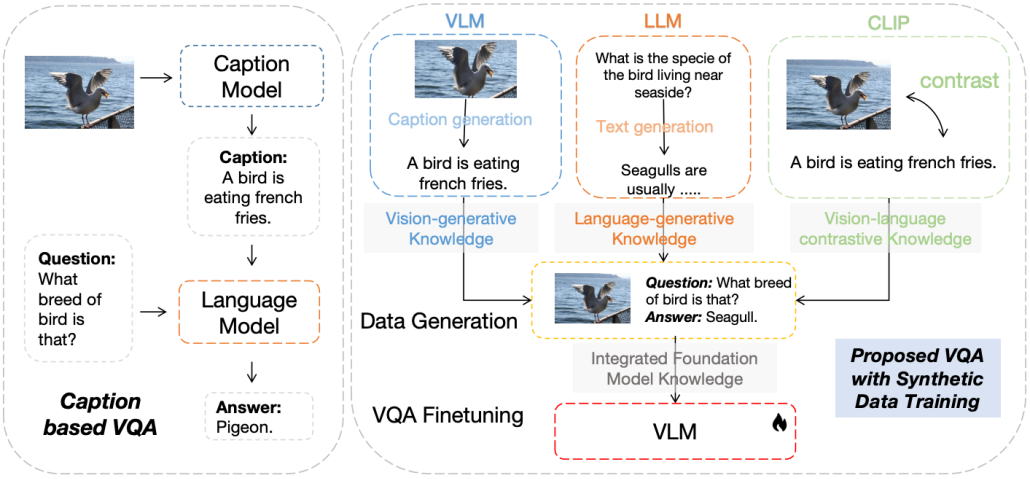


Figure 1: **Comparison of zero shot VQA paradigm.** Caption based VQA methods convert the image to captions and utilize language model to answer visual question. In contrast to previous approaches, we leverage foundation with diverse knowledge to generate synthetic data and finetune a generative vision language model to answer visual questions.

VQA called ***M***ultifaceted ***K***nowledge-guided ***D***ata ***G***eneration (**MKDG**). As shown in Figure 1, MKDG generates synthetic training data for VQA tasks by integrating multiple pre-trained vision and language models. Our key insights are as follows: 1) Generating synthetic data to transfer pre-trained model knowledge to a specific VQA domain is more efficient and effective than converting VQA to question answering tasks. 2) Constructing answer candidates with LLMs and using CLIP for data filtering enhances synthetic data quality which play a critical role in zero-shot VQA performance. 3) Converting VQA samples into the CLIP embedding space and using clustering-based data selection allows us to reduce the negative impact of data bias in synthetic data. Compared with widely used LLM synthetic data generation pipeline such as LLAVA (Liu et al., 2023b), we focus on design an efficient framework that automatically filter and select high-quality synthetic data.

Our MKDG framework integrates three pre-trained large models—CLIP (Radford et al., 2021), a generative vision-language model (VLM) (Li et al., 2023; Awadalla et al., 2023), and an LLM (Peng et al., 2023; Bai et al., 2023)—into a data generation and task-specific fine-tuning pipeline. Starting from a set of images, our pipeline consists of three main steps: 1) *Data generation with LLM and VLM.* We employ BLIP2 to generate captions for the unlabeled images to provide the LLM with visual information. An well-designed instruction and human annotated demonstrations are provided to prompt the LLM to create pseudo questions, answer options, and answers for an image. Here, the introduction of LLM knowledge alleviate the *deficiency in informativeness* in previous synthetic data generation based methods. 2) *Knowledge-based data Filtering with CLIP.* We assess the quality of the generated questions and answers using CLIP, filtering out data where CLIP and LLM make disparate choices, the image question pairs with low similarity and duplicated questions. The data selection reduce the incorrect answers in synthetic data by CLIP prior knowledge, thus reducing the *hallucination* of LLM. Furthermore, we utilize CLIP to map VQA samples into an embedding space and employ the K-means algorithm to cluster the generated data. We then select representative data points from these clusters to construct a high-quality synthetic dataset. 3) *Synthetic data training with VLM.* We finetune a moderate-sized VLM with the automatically curated synthetic dataset. This training paradigm directly processes the visual inputs and thus reduces the *visual information loss*, while the knowledge encoded in the synthetic dataset alleviates the *restricted reasoning ability* in the caption pre-trained VLM, such as BLIP2. Moreover, the fine-tuned VLM model achieves a much lower inference cost compared to previous LLM-based methods.

We empirically validate the MKDG framework on popular VQA benchmarks, OKVQA (Marino et al., 2019), A-OKVQA (Schwenk et al., 2022), VizWiz (Gurari et al., 2018), GQA (Hudson and Manning, 2019) and VQAV2 (Goyal et al., 2017), demonstrating it superior performance.

Our main contributions are summarized as following:

- We introduce MKDG, a framework that effectively transfer knowledge from large pretrained vision and language models to visual question answering via synthetic data training.
- We are the first to introduce prior knowledge in CLIP to perform automatic VQA data filtering and selection, which improve the data quality and reduce the negative impact of uneven data distribution in synthetic VQA data.
- Experimental results on popular VQA benchmarks demonstrate that the MKDG outperforms strong baselines equipped with larger GPT3 model.

## 2 Related Works

### 2.1 Large Pretrained Models

Recently, large language models (Brown et al., 2020a; Touvron et al., 2023a,b; Chung et al., 2022a) have achieved tremendous success. With a tremendous amount of training data and the scaling up of the number of parameters, large language models exhibit surprising capabilities, such as chain-of-thought reasoning (Wei et al., 2022), in-context learning (Brown et al., 2020b), and instruction following (Chung et al., 2022b). Vision-language models, including BLIP2 (Li et al., 2023), Flamingo (Alayrac et al., 2022) and BEIT-3 (Wang et al., 2023b), also benefit from large-scale training, which has led to significant progress in unified vision-language understanding and generation tasks. In particular, CLIP (Radford et al., 2021) employs a contrastive learning strategy on a huge amount of image-text pairs and shows impressive transferable ability over downstream tasks. CLIP provides an expressive vision-language joint embedding space, which enables content similarity measurement between image and text data.

### 2.2 Synthetic Visual Data Generation

With the development of pre-trained language models (Brown et al., 2020a; Chung et al., 2022a; Touvron et al., 2023b,a), automatic visual data generation attracts increasing interest. The early explorations (Changpinyo et al., 2022; Banerjee et al., 2020) utilize the T5-based question generation model and question answer model to convert caption datasets to synthetic VQA data. More recently, several works (Liu et al., 2023b,a; Wang et al., 2023a) generate visual instruction following datasets by instructing powerful LLM such as GPT4 (OpenAI, 2023a) to generate VQA data based on image annotations. In this work, we focus on the zero-shot VQA setting and propose a knowledge-guided data filtering strategy that can be seamlessly integrated into other synthetic data generation pipelines.

### 2.3 Zero-shot VQA

Visual Question Answering(VQA) involves advanced reasoning and image recognition. For zero-shot VQA, early methods (Changpinyo et al., 2022; Banerjee et al., 2020) generated synthetic VQA data from image captions using simple rules, then used this data for training. VQ2A (Changpinyo et al., 2022) extracted answer candidates from captions with rule-based methods and used a T5 (Raffel et al., 2019)-based question generator to create VQA samples, filtering the data with a T5-based QA model. Different from generating VQA samples from captions, recently proposed BLIP2 (Li et al., 2023) and Flamingo (Alayrac et al., 2022) train a vision-language model with caption data, which can generalize to VQA tasks without task-specific fine-tuning. With the advent of large pre-trained models, another promising strategy is to convert vision information into image captions and utilize the pre-trained large language model to comprehend the input and generate the expected answer. Recent works (Tsimpoukelli et al., 2021; Dai et al., 2022; Jin et al., 2022; Guo et al., 2023; Hu et al., 2022) utilize this strategy and obtain promising performance with powerful LLMs such as GPT3 (Brown et al., 2020a). In this work, we propose a synthetic data generation pipeline with novel knowledge-guided data filtering to address the zero-shot VQA task.

## 3 Methods

In this section, we introduce our MKDG framework, which incorporates CLIP (Radford et al., 2021), VLM (Li et al., 2023; Awadalla et al., 2023) and LLM (Peng et al., 2023; Bai et al., 2023) knowledge through synthetic VQA data training for zero-shot VQA. We depict the overall pipeline of our model in Section 3.1. In Section 3.2, we introduce the synthetic VQA data generation with LLM and Caption model. In Section 3.3, the knowledge-based data filtering is presented. Finally, we present the training with selected synthetic data in Section

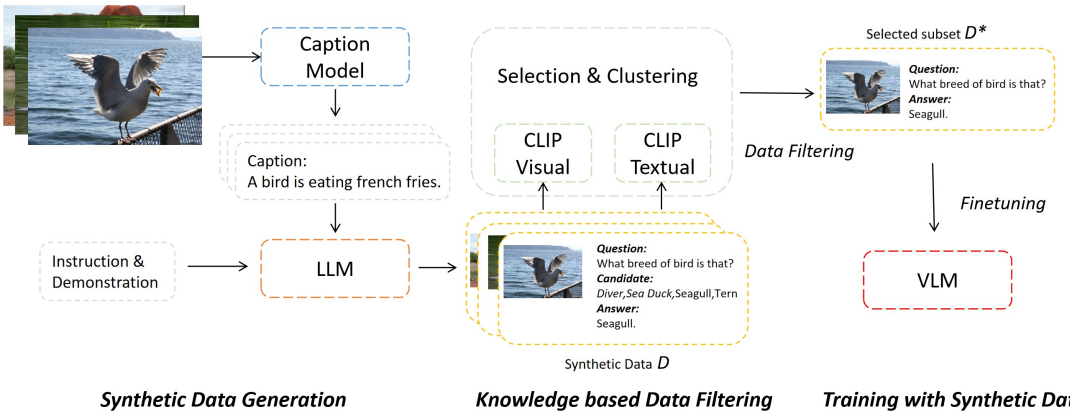


Figure 2: **The overall pipeline of our methods.** We utilize BLIP2 to convert the images to captions. The LLM is prompted to generate synthetic data based on caption, instruction and demonstration. We leverage CLIP to filter out low quality data and select a high quality subset with clustering strategy. The selected subset is used to finetune a BLIP2 for visual question answering.

3.4.

### 3.1 Overall Architecture

The primary challenge of zero-shot visual question answering (VQA) lies in effectively incorporating the knowledge encoded in these pre-trained models to benefit visual question answering. Our motivation is that synthetic data generation is the most efficient way to transfer Large Language Model (LLM) knowledge and we propose a synthetic VQA data generation pipeline. However, due to the visual information loss and bias in LLM, the hallucination and imbalanced data distribution in LLM generated data negatively impact the VQA data quality. To address the issue, we propose a knowledge-guided data filtering method that utilize the CLIP knowledge learned from contrastive vision-language pre-training to perform data filtering and selection.

The overall pipeline is illustrated in Figure 2. Specifically, we work with a set of unlabeled images $I$ and three pre-trained models, CLIP (Radford et al., 2021), a generative VLM (Li et al., 2023; Awadalla et al., 2023), and an LLM (Peng et al., 2023; Bai et al., 2023). In *Synthetic Data Generation* phase, we first use the VLM to generate a set of captions $C$ for each image. We then employ the LLM to generate questions $Q$, corresponding options $O$ and language model generated answers $A_{llm}$, which collectively form a synthetic dataset $D$. In *Knowledge Based Data Filtering* step, we leverage knowledge from CLIP to choose a high-quality subset $D^*$ from synthetic data $D$. The selected dataset $D^*$ is then employed to fine-tune the generative VLM, serving as our VQA model. We will introduce the details of each step in the subsequent sections.

### 3.2 Synthetic Data Generation

To transfer knowledge from pre-trained models to synthetic data, we start by constructing the synthetic data $D$ with LLM and BLIP. The construction pipeline contains the following steps: 1) We employ BLIP2 to transform unlabeled images $I = \{i^n | n = 1, \ldots, N\}$ into captions $C = \{c^n | n = 1, \ldots, N\}$ to provide visual information for LLMs where $N$ is the total number of images. 2) We write a task instruction and provide $8$ human annotated target VQA domain demonstrations as illustrated in Figure 3. The demonstrations are written based on target VQA domain question style to benefit the knowledge transfer process. More details are provided in the appendix. 3) The LLM follows these demonstrations to generate questions $Q = \{q^n | n = 1, \ldots, N\}$, options $O = \{o^n | n = 1, \ldots, N\}$ and answer $A_{llm} = \{a_l^n | i = 1, \ldots, N\}$. The overall process can be formulated as follows:

$$D = \{(i^n, q^n, a_l^n, o_1^n, \ldots, o_M^n) | n = 1, \ldots, N\} \quad (1)$$

where $M$ is the number of answer options for each sample. We clarify the necessity of generating answer options as follows: The generated answer options are the answers with highest confidence in the LLM question-answer generation space. We explicitly preserve this LLM knowledge instead of only keeping the top one answer. This strategy also forms the basis for discriminative models like CLIP to perform visual question answering during data selection.

In conclusion, we construct synthetic data $D$ from unlabeled images $I$, which is achieved by integrating the knowledge from LLM and BLIP2. However, the synthetic data $D$ is noisy and requires

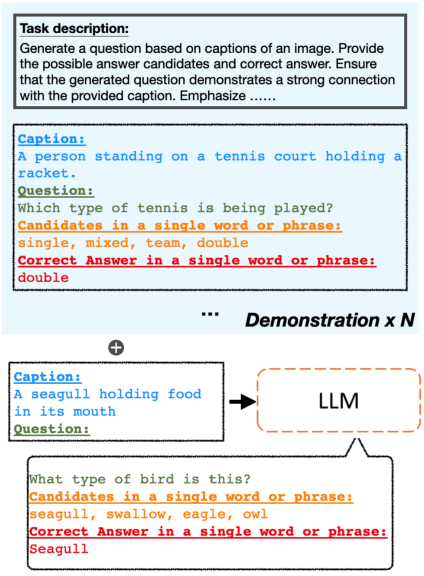


Figure 3: **Prompting in synthetic data generation.** We carefully designed 8 demonstrations and instructions to guide LLM to generate questions requiring outside knowledge to answer based on image caption.


Figure 4: **VQA data quality evaluation pipeline.** To evaluate the quality of VQA sample, we convert the question-answer pair to an image description and compute the cosine similarity between the image and the description.

data filtering.

### 3.3 Knowledge-guided Data Filtering

In the initial phases of experimental exploration, we identified two primary issues with the synthetic data $D$, which is generated by the LLM and caption model. The first issue is hallucination, where answers are incorrect or unrelated to the questions Secondly, generated questions are biased towards certain types, leading to duplication and creating low-quality data. We provide more evidence of this phenomenon in the appendix. To address the issues, we propose to perform VQA data quality evaluation and balanced data filtering using CLIP prior knowledge.

#### 3.3.1 Quality Evaluation of VQA Data

We evaluate the quality of synthetic data from two perspectives: *answer correctness* and *image relevance*. To improve the correctness of synthetic data, we compute the CLIP score for answer options $\{o_1^n, \ldots, o_M^n\}$ of each sample in the synthetic data $D$. The option with the highest CLIP score is regarded as the CLIP-chosen answer, denoted as $a_c^n$. We retain VQA samples where the *CLIP's answer* $a_c^n$ *and the LLM's answer* $a_l^n$ *are identical*, which mitigate hallucination caused by language model. To ensure the QA pair is relevant to the image, we utilize the CLIP score, which represents the relevance between the QA pair and the image in the CLIP embedding space. Specifically, we convert the question and answer into an image description, such as *"{question} Answer: {answer}"*, and calculate the cosine similarity between the image and the description. we discard samples with a CLIP score lower than a specified threshold. We illustrate the quality evaluation pipeline in Figure 4.

#### 3.3.2 Balanced Data Selection.

We illustrate the data selection pipeline in Figure 5. The skewed distribution of synthetic training data negatively impacts performance and hinders the effective transfer of pre-trained model knowledge via synthetic data $D$. We propose to use a clustering algorithm to select a balanced subset from synthetic data $D$. We firstly comprehensively represent VQA sample in an embedding space. Our proposed approach involves extracting question and image embeddings using CLIP and concatenating these embeddings to form the VQA embedding. The process is formulated as follows:

$$e_v^n = \text{CLIPVisual}(i^n), i^n \in I \quad (2)$$

$$e_t^n = \text{CLIPTextual}(q^n), q^n \in D \quad (3)$$

$$e^n = \text{Concat}(e_v^n, e_t^n) \quad (4)$$

where CLIPVisual and CLIPTextual indicate the visual and text encoders in CLIP. We denote the set of embedding of VQA samples by $E = \{e^n | n = 1, \ldots, N\}$. Given the high dimension of the VQA embedding $E$ and the limited scale of the available data, we employ Principal Component Analysis(PCA) to reduce the dimensionality of $E$ for efficient clustering. Subsequently, K-means is used to cluster the synthetic data $D$ based on $E$. In our prior investigations, we noted that the *CLIP score alone fails to accurately gauge the quality of VQA samples above a certain threshold*. To address this, we randomly sample an equal amount of VQA samples from each cluster, creating a balanced subset $D^* = \{(q^n, q^n, a^n) | n = 1, \ldots, N^*\}$, where $N^*$ is the total number of samples in selected datasets. This subset serves as the foundation for transferring knowledge from pre-trained models.

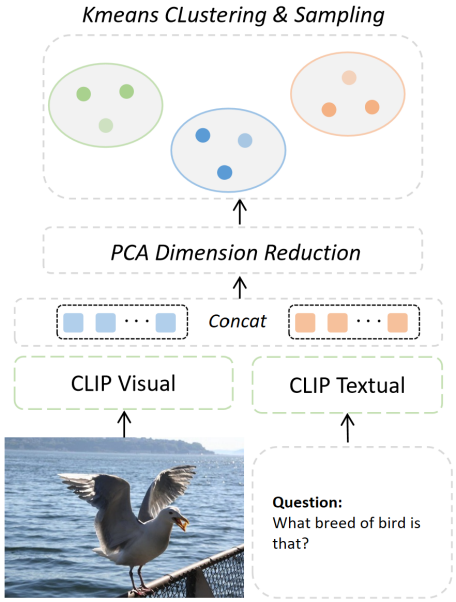


Figure 5: **The pipeline of data clustering.** The question and image in VQA sample are embedded by CLIP textual and visual encoders. We concatenate the resulting feature and utilize PCA to reduce the dimensionality. Finally, the VQA embeddings are clustered by Kmeans and we sample an equal size of data from each cluster.

### 3.4 Training with Synthetic Data

The selected synthetic data $D^*$ represent the collective intelligence of pre-trained models. We then finetune a pre-trained generative vision language model, such as BLIP2 (Li et al., 2023) and OpenFlamingo (Awadalla et al., 2023), using $D^*$ to excel in visual question answering.This pipeline not only harnesses the knowledge within pre-trained models but also learns to process visual input directly without relying on a caption, preventing visual information loss. Specifically, we train the VLM with the following objective:

$$L^n(\theta) = -\frac{1}{l_n}\sum_{j=1}^{l_n} \log P_\theta(w_j|w_{<j}, i^n, q^n) \quad (5)$$

where $l_n$ is the number of words in $a^n$, $w_j$ is the $j$-th word in answer text $a^n$ and $i^n$, $q^n$ are the image and question in the selected synthetic data $D^*$. $\theta$ are the parameters of the vision language model for fine-tuning. In summary, our approach involves training a generative vision language model, incorporating knowledge learned from multiple pre-trained models to excel in visual question answering.

## 4 Experiments

In this section, we demonstrate MKDG's effectiveness on zero-shot VQA task. We show that our method achieves promising performance on VQA benchmarks without VQA annotation. Then we conduct ablation experiments on the contribution of each component in our pipeline.

### 4.1 Experimental Setting

#### 4.1.1 Datasets and Evaluation

Our experimental evaluations are performed on five benchmark datasets for knowledge-based visual question answering: OK-VQA (Marino et al., 2019), A-OKVQA (Schwenk et al., 2022), VizWiz (Gurari et al., 2018), GQA (Hudson and Manning, 2019), and VQAV2 (Goyal et al., 2017). OK-VQA contains 14K image-question pairs focused on open-ended questions with detailed answers. A-OKVQA improves on OK-VQA with 25K pairs of higher-quality questions and images. VizWiz features diverse images and questions from visually impaired individuals, with a validation set of about 4K images. GQA tests visual reasoning with 5K images in the test-dev set. VQAV2 is a general benchmark with 200K images in the validation set. To align with prior work (Tiong et al., 2022; Guo et al., 2023; Du et al., 2023), we measure standard VQA Accuracy (Goyal et al., 2017) on OK-VQA, A-OKVQA, and VQAV2, and top-1 accuracy for VizWiz and GQA. For all benchmarks, we use an open-ended generation strategy. Following (Li et al., 2023; Awadalla et al., 2023), we use beam search with a size of 5 and a length penalty of -1 to ensure answers fit the desired format.

#### 4.1.2 Implementation Detail

To generate the synthetic data and serve as the outside knowledge source for our pipeline, we use Vicuna-7B (Peng et al., 2023) and QWEN72B (Bai et al., 2023) as the language model. Vicuna is finetuned with the conversation collected from GPT4 (OpenAI, 2023b) based on LLAMA2 (Touvron et al., 2023b), which serves as a powerful and feasible language model for research purposes. QWEN72B is an open-source language model series that encompasses distinct models with varying parameter counts. For the generative vision language model, we adopts BLIP2 (Li et al., 2023) with OPT-2.7B(Zhang et al., 2022), FlanT5 XL (Chung et al., 2022a) and OpenFlamingo3B (Awadalla et al., 2023) for our experi-

| Model | Infer Param. | OK-VQA | A-OKVQA | VizWiz | GQA | VQAV2 |
|---|---|---|---|---|---|---|
| **In-context-learning** | | | | | | |
| PICa & 16 Demo (Yang et al., 2022) | 175B | 46.9 | - | - | - | - |
| PromptCap-FlanT5 XXL & 32 Demo (Hu et al., 2022) | 11B | 42.0 | - | - | - | - |
| PromptCap & 1 Demo (Hu et al., 2022) | 175B | 48.7 | - | - | - | - |
| **Zero-shot VQA** | | | | | | |
| Frozon (Tsimpoukelli et al., 2021) | 7B | 5.9 | - | - | - | 29.5 |
| VLKD (Dai et al., 2022) | <1B | 13.3 | - | - | 29.3 | 42.6 |
| FewVLM (Jin et al., 2022) | <1B | 16.5 | - | - | - | - |
| PICa (Yang et al., 2022) | 175B | 17.7 | - | - | - | - |
| VQ2A† (Changpinyo et al., 2022) | - | 19.8 | - | - | **50.0** | 57.9 |
| PNP-VQA3B (Tiong et al., 2022) | 3B | 34.1 | 42.3 | - | - | 62.1 |
| PNP-VQA11B (Tiong et al., 2022) | 11B | 35.9 | 41.9 | - | - | **64.8** |
| LAMOC11B (Du et al., 2023) | 11.4B | 40.3 | 37.9 | - | - | - |
| openFlamingo(9B) (Awadalla et al., 2023) | 9B | 37.8 | - | 27.5 | - | 52.7 |
| Flamingo (3B) (Alayrac et al., 2022) | 3B | 41.2 | - | 28.9 | - | 49.2 |
| Flamingo (9B) (Alayrac et al., 2022) | 9B | 44.7 | - | 28.8 | - | 51.8 |
| Img2LLM6.7B (Guo et al., 2023) | 6.7B | 38.2 | 33.3 | - | - | - |
| Img2LLM-175B (Guo et al., 2023) | 175B | 45.6 | 42.9 | - | - | 60.6 |
| openFlamingo(3B) | 3B | 28.2 | 26.2 | 23.7 | - | 44.6 |
| openFlamingo(3B) + MKDG | 3B | 39.7(+11.5) | 36.4(+11.5) | 42.8(+19.1) | - | 50.0(+5.4) |
| BLIP-2 OPT2.7B | 3.1B | 30.2 | 26.3 | 14.3 | 33.9 | 50.1 |
| BLIP-2 OPT2.7B + MKDG | 3.1B | **48.3**(+18.1) | 42.8(+16.5) | 43.7(+29.4) | 41.3(+8.3) | 57.1(+7.0) |
| BLIP-2 FlanT5(XL) | 7.8B | 39.4 | 40.0 | 17.9 | 44.2 | 62.6 |
| BLIP-2 FlanT5XL + MKDG | 3.4B | 46.3(+6.9) | **46.0**(+6.0) | **46.0**(+28.1) | 42.5(-1.7) | 60.3(-2.3) |

Table 1: **Comparison with state-of-the-art methods on popular VQA benchmarks.** We present the inference time parameters size in *Infer Param.*. The *In-context learning* indicate methods utilize demonstrations and we mark the number of used demonstrations after &. The † indicates the inference strategy is classification instead of open-word generation.

ments. We use the CLIP (Radford et al., 2021) with ViT-L/14 visual encoder to select high-quality data. For synthetic data generation, we use about 80k images from COCO2014 training set (Lin et al., 2014), and the captions are generated by the BLIP2 OPT-2.7B. The number of anwer options $M$ is set to 4 in our experiments. Before knowledge-guided data filtering, we collect about 230k valid VQA samples from QWEN72B and about 310k valid VQA samples from Vicuna7B. We combined data generated by two different language models to integrate the knowledge from both. For Vicuna 7B, it takes about one day to generate data with 4 NVIDIA A40 GPUs and we use API to access Qwen72B. For K-means clustering, we use PCA (Pearson, 1901) to reduce the dimension to 256 and the number of clustering centers is set to 400. We sample 40k VQA samples for the model training. We provide a hyper-parameter sensitive analysis in the Appendix. We generate individual synthetic datasets for each benchmark except the A-OKVQA and OK-VQA, which share a synthetic dataset. During model training, we only finetune the Q-former in BLIP2 and adopt the text-aware visual feature extraction. For OpenFlamingo3B, we follow the pre-trained training setting, which finetunes the perceiver and gated dense cross-attention layers. The learning rate is set to 1e-4 with a cosine annealing AdamW optimizer (Loshchilov and Hutter, 2017). The training takes about 1h on 4 NVIDIA A40 GPUs.

#### 4.1.3 Baselines

We compare MKDG with strong baselines in zero-shot VQA task to showcase the effectiveness. The baselines can be summarized into two categories, corresponding to the two sections in Table 1. *LLM with In-context-learning* consists of methods (Yang et al., 2022; Hu et al., 2022) that transfer images into captions and then use a pre-trained large language model to accomplish VQA task. Such methods rely on in-context learning for better performance and the demonstrations come from the training set of the target dataset, which introduces additional target domain information in inference. In *Zero-shot VQA*, methods (Li et al., 2023; Alayrac et al., 2022; Tsimpoukelli et al., 2021; Dai et al., 2022; Jin et al., 2022) leverage large-scale caption data for pretraining, thus obtaining generalizable ability to accomplish zero-shot VQA task. VQ2A (Changpinyo et al., 2022) utilizes a rule-based strategy to extract answer candidates from the caption and use FlanT5 XXL to generate synthetic data. Methods (Tiong et al., 2022; Du et al., 2023; Guo et al., 2023) focus on providing the language model with a better caption to achieve better visual question performance.

### 4.2 Zero-shot Visual Question Answering

#### 4.2.1 Knowledge Based VQA

To comprehensively evaluate our method, we compare MKDG with baselines on knowledge-based VQA benchmarks, including OK-VQA test set, A-OKVQA validation set, and Vizwiz validation set, as shown in Table 1. MKDG shows significant performance improvements over previous methods on OK-VQA and A-OKVQA, achieving results comparable to GPT-3 175B-based methods like PICa (Yang et al., 2022), Img2LLM-175B (Guo et al., 2023), and PromptCap (Hu et al., 2022), but with only 1.7% of the inference parameters. On the more challenging Vizwiz benchmark, MKDG outperforms zero-shot BLIP2 (Li et al., 2023) and OpenFlamingo (Awadalla et al., 2023), demonstrating its adaptability to previously underperforming task domains. Compared to caption-based methods (Tiong et al., 2022; Guo et al., 2023; Du et al., 2023) without in-context learning, our approach achieves state-of-the-art performance with smaller pre-trained models. This highlights the efficiency of our knowledge transfer paradigm and the detrimental impact of information loss when converting images to captions for VQA performance.

#### 4.2.2 Results on VQAV2 and GQA

Additionally, we evaluate MKDG on general-purpose VQA, VQAV2 (Goyal et al., 2017) and relational reasoning VQA, GQA (Hudson and Manning, 2019). As shown in Table 1, though MKDG brings improvements over pre-trained BLIP2 OPT2.7B, the performance in BLIP2 FLan5XL decreases slightly compared with the pre-trained one, which already achieves strong performance. Our approach aims to enhance the model's performance to a moderate level in the new VQA domain where its performance has been sub-optimal.

### 4.3 Ablation Study

#### 4.3.1 Data Selection Strategy

In this section, we conduct an ablation study to provide a comprehensive interpretation of our proposed methods. We disentangle our knowledge-based data selection strategy and select multiple datasets with different strategies. In detail, ***Base*** indicate that no data filtering is applied and we preserve 540k noisy synthetic data. The ***Ranking*** means that we only preserve the top k samples sorted by the CLIP score. In ***Consistent***, we remove the VQA samples where the language model's answer $a_l^i$ and CLIP answer $a_c^i$ are inconsistent. The ***Clustering*** indicate that the VQA samples below a CLIP score threshold are removed. For the rest of the samples, we perform K-means clustering and randomly sample an equal size of data from each cluster. our methods choose the ***Consistent + Clustering*** and we present the OK-VQA, A-OKVQA performance of these strategies in Table 2. The result demonstrates that our data selection is the most effective way of extracting knowledge from synthetic data. Additionally, we observed that the ***Consistent*** outperform ***Consistent + Ranking*** on OK-VQA, which proves that the high CLIP scores don't exhibit a positive correlation with high VQA data quality. However, as ***Ranking*** largely outperforms ***Base***, we conclude that CLIP scores are capable of filtering out low-quality VQA samples. These findings contribute to our final strategy ***Consistent + Clustering***.

| Strategy | Total Data | OK-VQA | A-OKVQA |
|---|---|---|---|
| Base | 540k | 43.6 | 37.8 |
| Ranking | 540k | 43.7 | 38.1 |
| Consistent | 211k | 47.2 | 42.1 |
| Ranking + Consistent | 211k | 47.0 | 41.0 |
| Consistent + Clustering | 211k | **48.3** | **42.8** |

Table 2: **Ablation on data selection strategy.** For a fair comparison, we train a BLIP2-OPT2.7b model with data selected by different strategies. Total Data indicate the available data size after filtering and we sample 40,000 data from available data to train the model.

## 5 Conclusion

We propose MKDG, a zero-shot Visual Question Answering (VQA) framework that leverages pre-trained model knowledge through synthetic data generation. MKDG encodes the extensive knowledge of large language models (LLMs) in synthetic data and uses the vision-language knowledge in CLIP to filter out noise VQA samples. Specifically, MKDG employs a caption model to provide visual information to the LLM, prompting it to generate synthetic VQA data. To mitigate hallucinations and uneven data distribution, we use CLIP's prior knowledge to filter out incorrect VQA data and select a high-quality subset through clustering. Finally, we train a moderate-sized generative vision-language model with the curated data, integrating the knowledge from CLIP, LLM, and VLM. Experimental results demonstrate the superior performance of our method across various popular VQA benchmarks.

## 6 Limitation

**Limited Improvements in General VQA.** Despite the promising results on knowledge-based VQA benchmarks, our approach has several limitations. Specifically, because our generated data rely on integrating foundation model knowledge, MKDG does not achieve significant improvements on general VQA benchmarks such as VQAV2 and GQA, where questions focus on visual perception and scene understanding. Additionally, existing models like BLIP2 (Li et al., 2023) and OpenFlamingo (Awadalla et al., 2023) already perform well on these benchmarks. Therefore, our paradigm is more suited for facilitating knowledge transfer within sub-optimal target domains.

**Sub-optimal Domain Generalization Ability.** The proposed MKDG pipeline generates synthetic datasets for specific VQA domains by prompting LLMs with domain-specific prompts. While it achieves significant improvements in the target VQA domain, the fine-tuned model's performance gains in other VQA domains are limited. For example, the synthetic VQA dataset is generate based on the prompts regarding OK-VQA benchmark, the fine-tuned model on this datasets may not benefit the performance on GQA benchmark. Therefore, our pipeline serves as an efficient framework for VQA domain adaptation without requiring annotations.

## References


Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak 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. *ArXiv*, abs/2308.01390.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenhang Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, K. Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Yu Bowen, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xing Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *ArXiv*, abs/2309.16609.

Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, and Chitta Baral. 2020. Weaqa: Weak supervision via captions for visual question answering. *arXiv preprint arXiv:2012.02356*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020a. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020b. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Soravit Changpinyo, Doron Kukliansky, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. 2022. All you may need for vqa are image captions. *arXiv preprint arXiv:2205.01883*.

Lin Chen, Jinsong Li, Xiao wen Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multi-modal models with better captions. *ArXiv*, abs/2311.12793.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022a. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022b. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Enabling multimodal generation on CLIP via vision-language knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2383–2395, Dublin, Ireland. Association for Computational Linguistics.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Yifan Du, Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Zero-shot visual question answering with language model feedback. *arXiv preprint arXiv:2305.17006*.

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)*.

Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. 2023. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10867–10877.

Danna Gurari, Qing Li, Abigale Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3608–3617.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2022. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702.

Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. 2022. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. *Preprint*, arXiv:2110.08484.

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

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.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *ArXiv*, abs/2304.08485.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 852–869. Springer.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

OpenAI. 2023a. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

OpenAI. 2023b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Karl Pearson. 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

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.

Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven CH Hoi. 2022. Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training. *arXiv preprint arXiv:2210.08773*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. 2023a. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *ArXiv*, abs/2311.07574.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2023b. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

# A Appendix

### A.0.1 Visualization.

This section aims to provide a visual depiction of the characteristics and quality of the synthetic data, thereby facilitating a deeper comprehension of our methodology. Presented in Figure 6, we showcase three examples illustrating the generated synthetic data alongside the predictions made by both a large language model and CLIP. Notably, in the first two cases, both the large language model and CLIP exhibit concurrence, while in the final case, discrepancies arise between their choices. When both large language model and CLIP converge in agreement, it signifies a robust alignment between the generated questions, answers, and corresponding images. This alignment implies that the synthetic data tends to possess favorable attributes: answers are predominantly correct, questions maintain relevancy to the image content, and there is less ambiguity in answers.

### A.0.2 Ablation Study of language model and caption quality.

To provide more insights into the role of language model and caption quality in MKDG framework, we finetune the models with synthetic data generated by different language model and caption. The language model serve as the knowledge source for synthetic data while the captions are source of visual information for language model. To explore the importance of caption quality in VQA data generation, we propose three categories of captions to conduct experiments. The *BLIP2 OPT2.7B* indicates the short and relatively low quality captions generated by the pre-trained BLIP2 OPT2.7B. The *ShareGPT4V* indicates the dense caption generated by ShareGPT4V (Chen et al., 2023), which contained a GPT4V (Chen et al., 2023) style detailed image description. The *Ground Truth* indicates the precise caption annotated by human. The results are shown in Table 3, we observe the quality of data generated by QWEN72B is higher than that of Vicuna7B when the caption is generated by ShareGPT4V or ground truth caption. However, Vicuna 7B generates better data with BLIP2 generated caption. The performance gap demonstrates that ***scaling up language model will not guarantee improvements*** under MKDG framework. Another observation is that *Ground Truth* captions achieve the best results across all language models, which reveal that the ***precision of caption is the most important factor*** in MKDG. Furthermore, we utilize the mixture of data generated by the two LLMs and achieve highest performance. This phenomenon indicates the ***language model contains different inductive bias*** in generating VQA data. Our proposed clustering and selection strategies extract and fuse the informative parts of the data generated by two LLMs thus achieving better performance.

| | BLIP2 OPT2.7B | ShareGPT4V | Ground Truth |
|---|---|---|---|
| Vicuna7B | 46.8 | 43.9 | **47.5** |
| Qwen72B | 44.1 | 46.1 | **47.6** |
| Mixed Data | 48.3 | 47.2 | **48.4** |

Table 3: **The performance with different language model and caption.** For fair comparison, we train a BLIP2 OPT2.7B with the data generated by different language model and captions. The results on OKVQA are reported in the table.

## A.1 Bias in Generated Questions

We mentioned that the generated questions from large language model are biased towards certain question types in main paper section *Knowledge Based Data Filtering*. In order to provide a clear understanding of this phenomenon, We calculated the frequency of the first four words in the questions from all generated data and present the **top 10** question type in Figure 7. We observe that the top one question type *What is the name* account for *22.94%* data and the sum of top 4 question type account for *49.8%* data. The statistical results of this analysis demonstrate that the generated questions exhibit a bias towards certain question types, which leads to duplication in data and harms the VQA performance as demonstrated in main paper ablation study Table 2.

## A.2 Inference Time Comparison.

For a quantitative comparison of inference efficiency, we measured the inference time of BLIP2-OPT2.7B and Vicuna 7B/13B on an A40 GPU, averaging the inference time over 1000 samples. In the case of Vicuna, we utilized 8 demonstrations to guide answer generation. BLIP2 achieved an inference time of **1.009** seconds per sample. In contrast, Vicuna 7B exhibited an inference time of **4.893** seconds per sample, while the 13B model showed **10.101** seconds per sample. The inference time of our methods is equal to BLIP2-OPT2.7B, which indicates that our methods significantly improved inference efficiency compared to caption-based VQA methods and achieve strong performance.

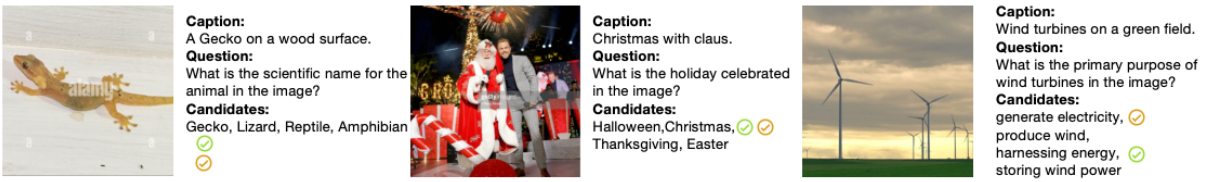


Figure 6: **Visualization of the synthetic data** We show the example caption, question, candidates and answer generated in our pipeline. We use red check mark to indicate the answer chosen by large language model and the green check mark to indicate the answer chosen by CLIP.

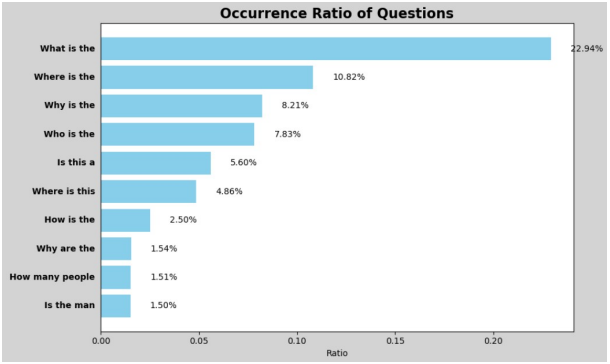


Figure 7: **The question type distribution in generated data.** We present percentage of the top 10 question type in generated data.

| PCA Dimension | 32 | 64 | 128 | 256 | 512 |
|---|---|---|---|---|---|
| vqa score | 47.67 | 48.38 | 48.26 | 48.52 | 47.77 |
| Number of Clusters | 200 | 300 | 400 | 500 | 600 |
| vqa score | 48.80 | 48.12 | 49.01 | 48.78 | 48.84 |

Table 4: Hyperparameters ablation experiments.

### A.3 Analysis on Hyperparameters.

We performed ablation study on hyperparameters on the OK-VQA training set with varied choices for PCA reduction dimension and the number of Kmeans clusters. The results are presented in Table 4. Notably, our methods is insensitivity to changes in PCA reduction dimension and the number of Kmeans clusters.

### A.4 Instruction and Demonstration

The instruction and demonstration are critical for guiding the LLM to generate desired data format, which is especially important in knowledge based VQA benchmarks where the answers are usually one word. To ensure the diversity of generated question types, we introduce question prefixes such as *['What', 'How', 'Where', 'Who', 'Why', 'Is']*.We provide the complete instruction for OK-VQA dataset as follow: *Generate a question based on captions of an image. Provide the possible answer candidates and correct answer. Ensure that the generated question demonstrates a strong connection with the provided caption. Emphasize that the question should be informative and require knowledge within the LLM to answer. For instance, instruct the model to inquire about specific details mentioned in the caption, demanding comprehension of external knowledge to respond accurately.*

For the demonstrations, we manually write 8 OK-VQA style VQA sample and use the corresponding captions from MSCOCO. To get the answer candidates, we use ChatGPT to generate the reasonable answer candidates. The complete demonstration are presented as follow:

*Caption: A large white, yellow and red bus driving down a street. A white, red and yellow transit bus is making its way through a town. A red, white and yellow bus on a street. A Victory Liner bus driving down a street. A red, yellow and white transit bus travelling down a street. Question: Is this a privately or publically owned vehicle? Candidates in a single word or phrase: public private government commercial Correct Answer in a single word or phrase: private*

*Caption: A person standing on a tennis court holding a racket. A person holding a tennis racket at a tennis court. a little kid that has a racket in his hand. A young man holding a tennis racquet on top of a tennis court. There is a group of people playing tennis on a court. Question: Which type of tennis is being played? Candidates in a single word or phrase: single mixed team double Correct Answer in a single word or phrase: double*

*Caption: A collection of pictures showing the before and after of a bathroom remodel. a bathroom*

*slowly getting remodeld with different pics. Three different photos of a bathroom being remodeled. The room was remodeled and the bathtub was removed. Three images of the process of a bathroom remodel. Question: What material is the bathtub made out of? Candidates in a single word or phrase: porcelain marble ceramic acrylic Correct Answer in a single word or phrase: ceramic*

*Caption: A brown horse standing next to a building wearing a blanket. A horse inside of a barn getting a bath. A horse tied to a stable wearing a pink and blue blanket. a horse wearing something tethered to a wall. A horse stands near the stalls wearing a blanket. Question: Where is this photo taken? Candidates in a single word or phrase: barn stable farm ranch Correct Answer in a single word or phrase: stable*

*Caption: A group of livestock are grazing in bright green grass. A group of dogs are roaming around a bright green field. A green pasture with cattle spread around it. Cows and horses graze in a wide open green field. Cattle and horses grazing in a green pasture. Question: Are these different animals in this picture or all they all the same animal? Candidates in a single word or phrase: same different mixed varied Correct Answer in a single word or phrase: different*

*Caption: Beach umbrellas provide shade at the beach as people walk the shoreline. People at the beach with several umbrellas scattered around. A beach scene with several colorful bathers umbrellas. People spending time on a beach during the summer. some people at a beach with rainbow colored umbrellas. Question: Who invented the colorful objects in the image? Candidates in a single word or phrase: samuel fox mary anderson george sage john w. dickinson Correct Answer in a single word or phrase: samuel fox*

*Caption: Two teams' coaches shake hands on a baseball field. A picture of three people talking to each other. A man in black pants and a white shirt holds a baseball and a man in a baseball uniform stands next to a man with sunglasses and a blue t-shirt. Baseball player and manager meeting before the game. A friendly chat on the field at a baseball game. Question: When you fill a glass to the top you are also referring to which part of the headgear worn here? Candidates in a single word or phrase: brim crown visor strap Correct Answer in a single word or phrase: brim*

*Caption: A two-person vanity is below a mirror in the bathroom. A double-sink vanity is in front of a wide mirror with side lighting in this rest room. An elegant bathroom has a light up mirror, marble counter tops and dual sinks. A nice marble tile sink with his and her. A lighted mirror illuminates two tidy bathroom sinks. Question: What is that counter top made of? Candidates in a single word or phrase: granite quartz marble ceramic Correct Answer in a single word or phrase: marble*