

Visual Text Matters: Improving Text-KVQA with Visual Text Entity Knowledge-aware Large Multimodal Assistant

Abhirama Subramanyam Penamakuri and Anand Mishra

Indian Institute of Technology Jodhpur

{penamakuri.1,mishra}@iitj.ac.in

<https://vl2g.github.io/projects/LMM4Text-KVQA/>

Abstract

We revisit knowledge-aware text-based visual question answering, also known as Text-KVQA in the light of modern advancements in large multimodal models (LMMs), and make the following contributions: (i) We propose VisTEL – a principled approach to perform visual text entity linking. The proposed VisTEL module harnesses a state-of-the-art visual text recognition engine and the power of a large multimodal model to jointly reason using textual and visual context obtained using surrounding cues in the image to link the visual text entity to the correct knowledge base entity. (ii) We present KALMA – knowledge-aware large multimodal assistant that augments an LMM with knowledge associated with visual text entity in the image to arrive at an accurate answer. Further, we provide a comprehensive experimental analysis and comparison of our approach with traditional visual question answering, pre-large multimodal models, and large multimodal models, as well as prior top-performing approaches. Averaging over three splits of Text-KVQA, our proposed approach surpasses the previous best approach by a substantial 23.3% on an absolute scale and establishes a new state of the art. We make our implementation publicly available.

1 Introduction

In the past few years, the research community has shown significant interest in visual question answering based on text appearing in images, as evidenced by the emergence of OCR-VQA (Mishra et al., 2019), ST-VQA (Biten et al., 2019b) and TextVQA (Singh et al., 2019b). Giving another aspect to these problems by leveraging external knowledge for text-based visual question answering, (Singh et al., 2019a) introduced a task called Text-KVQA. The Text-KVQA presents a unique challenge: given an image containing textual entities like business brands, book titles, or movie titles, the task is to answer questions that require

external knowledge about these entities. Addressing Text-KVQA involves detecting text in images, recognizing it, linking it to a knowledge base, and employing visual context and knowledge base for reasoning to provide an answer. Since the introduction of this problem, several advancements have happened in visual text understanding as well as vision and language models. In this work, we revisit Text-KVQA by leveraging these modern advancements and propose a framework that judiciously integrates various components of contemporary architecture.

The emergence of large multimodal models (LMMs)¹ represents a significant trend in the literature on vision and language (Zhang et al., 2022; Chung et al., 2024; Touvron et al., 2023; Liu et al., 2024; Zhu et al., 2023; Ye et al., 2023; Penedo et al., 2024; Ouyang et al., 2022). Over the past few years, many large-scale language and vision models have been developed, demonstrating exceptional performance across various tasks including, but not limited to, image captioning, visual question answering, multimodal reasoning, and visual grounding. We believe that pretrained LMMs hold great potential for addressing Text-KVQA. These models are rich in the implicit knowledge learned by large-scale pretraining. However, despite their numerous advantages, they are not without drawbacks, notably hallucinations. This challenge becomes particularly apparent in Text-KVQA, where precise reasoning about entities depicted in images and associated knowledge is required. Consider the following scenario where a customer, after finishing their meal at a restaurant store, takes a picture of the store signboard and enquires about a possible future online delivery, asking, ‘Where can I place an online order from this store?’ (Figure 1(a)). Existing LMMs often hallucinate over

¹We refer to both large multimodal model and large vision and language models as LMM in this work.

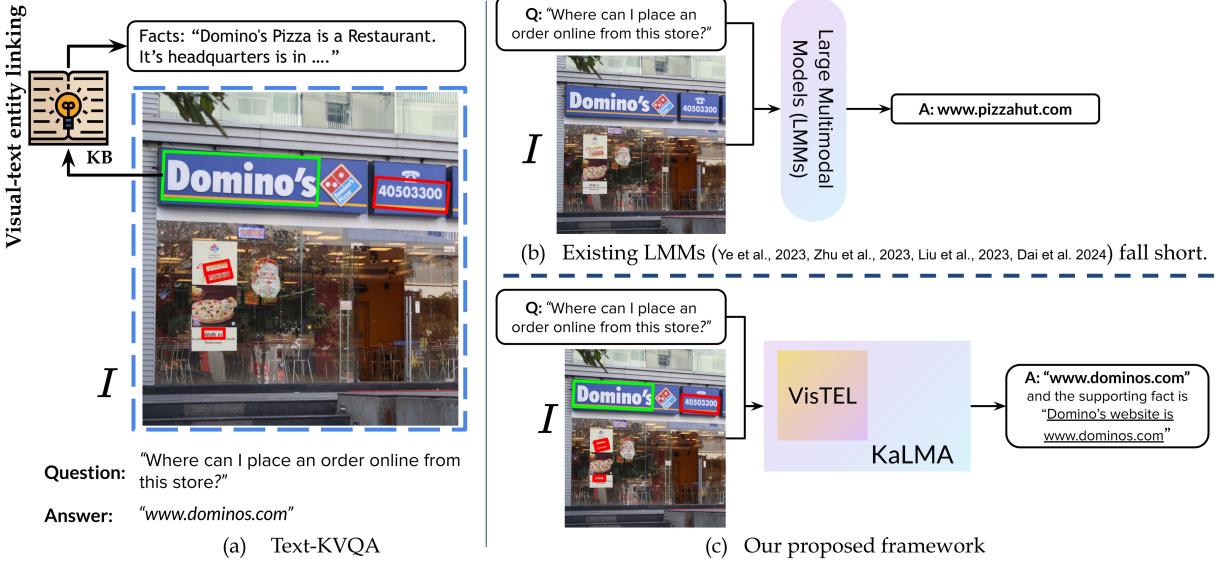


Figure 1: (a) **Text-KVQA** (Singh et al., 2019a): Given an image containing a named entity as visual text, e.g., “Domino’s” in this illustration, the aim is to answer the question by leveraging explicit knowledge about the visual text entity. (b) Large Multimodal Models are one obvious choice for solving such tasks today. However, they alone are insufficient as they hallucinate on visual objects. (c) We propose a novel approach – KaLMA that augments an LMM with specialized visual text recognition and retrieved relevant knowledge obtained using visual text entity linking by proposed VisTEL. Our approach establishes a new state-of-the-art for this task.

the pizza present in the image and points to the website of ‘Pizza Hut’ instead of ‘Domino’s’ (Figure 1(b)); whereas complementing the LMM with an explicit visual text entity linking followed by knowledge-retrieval helps overcome hallucination (Figure 1(c)), thereby generating an accurate answer to the given question. Our model is developed on this hypothesis.

We address Text-KVQA by introducing an architecture, namely KaLMA – knowledge-aware large multimodal assistant that first invokes our proposed visual text entity linker or VisTEL – an LMM-architecture that links visual text entities to the associated knowledge base (Illustrated in Figure 1 (c)). Once the entities are linked to the knowledge base, the associated knowledge is retrieved and augmented to a large multimodal model to answer visual questions.

To summarize, our contributions are as follows: (i) We revisit Text-KVQA – a task originally introduced by (Singh et al., 2019a) in the light of the latest advancements in large multimodal models. To this end, we benchmark latest LMMs on Text-KVQA. Our study highlights that LMMs although powerful, often ignore visual text present in the images, resulting in hallucinations.

(ii) We propose a principled approach called VisTEL for linking visual text entities that appear in images to a knowledge base. VisTEL is an LMM-based architecture that leverages the sur-

rounding OCR-extracted texts obtained using a specialized text recognition module and the visual context within the image to perform highly accurate entity linking for visual text entities. (iii) We introduce KaLMA – a Knowledge-aware Large Multimodal Assistant, which enhances a large-multimodal model, specifically LLaVA (Liu et al., 2024) by integrating retrieved knowledge from our proposed VisTEL. This augmentation facilitates robust vision and language reasoning, thereby enabling superior knowledge-aware text-based visual question answering. (iv) We conduct extensive experiments and ablation to show the superior performance of our proposed framework over competitive approaches and state of the art. We provide several exciting insights about our design choice, attribution ability of KaLMA, and addressing hallucination issues of LMMs. Our proposed approach advances state of the art on Text-KVQA by 18.2% on scene, 19.6% on book covers, and 32.2% on movie poster splits of the dataset on an absolute scale.

2 Related Work

KVQA Tasks: Visual Question Answering is a well-studied task (Antol et al., 2015; Goyal et al., 2017). This task has been extended to scenarios that require the ability to read text within images, leading to the development benchmarks such as

ST-VQA (Biten et al., 2019b,a), TextVQA (Singh et al., 2019b), DocVQA (Mathew et al., 2021), and OCR-VQA (Mishra et al., 2019). While these benchmarks were successful in their intent of integrating reading and reasoning abilities in VQA, they are often restricted to reasoning around what is visually apparent. To address this gap and encourage models to perform reasoning beyond visually apparent facts, (Singh et al., 2019a) introduced knowledge-aware Text-based VQA task. Distinctively different from other knowledge-aware visual question answering tasks such as KB-VQA (Wang et al., 2017b), FVQA (Wang et al., 2017a), KVQA (Shah et al., 2019), OK-VQA (Marino et al., 2019), and Infoseek (Chen et al., 2023), Text-KVQA deals with reasoning over visual text entities and associated knowledge to arrive at answer.

Methods Prior to Large Multimodal Models:

Early methods to solve knowledge-aware VQA tasks focus on leveraging knowledge in the form of triplets (Narasimhan et al., 2018; Narasimhan and Schwing, 2018; Wu et al., 2016), or sub-knowledge-graph (Zhang et al., 2018; Singh et al., 2019a) or memory facts (Weston et al., 2015). Later, transformer architectures (Vaswani et al., 2017) owing to their ability to encode intrinsic knowledge using large-scale pretraining, have become defacto for addressing KVQA.

Inspired by the hybrid models, e.g. (Lewis et al., 2020; Guu et al., 2020) where intrinsic knowledge of transformer architectures is complemented with explicit external knowledge; researchers proposed hybrid methods such as ConceptBERT (Gardères et al., 2020), KRISP (Marino et al., 2021), and REVEAL (Hu et al., 2023b) which augment the multimodal transformers with explicitly retrieved external knowledge.

Emergence of Large Multimodal Models: The early success of large-scale pretraining on the downstream tasks demonstrated by the foundation models, e.g., BERT (Devlin et al., 2019) and GPT (Radford et al., 2019) paved the way for the researchers to scale the model and the data used for pretraining. GPT-3 (Brown et al., 2020) is an early large language model (LLM) demonstrating reliable performance on many downstream tasks. Following this, several LLM variants (Zhang et al., 2022; Chung et al., 2024; Workshop et al., 2022; Penedo et al., 2024; Touvron et al., 2023) have been introduced. Researchers adopted these LLMs to vision-language research, with the key idea being aligning the visual information with the linguistic

information of the LLMs to come up with large multimodal models (LMMs) (Tsimpoukelli et al., 2021; Penedo et al., 2024; Li et al., 2023; Zhu et al., 2023; Ye et al., 2023; Liu et al., 2024). Recently, LMMs have become first-hand solutions for many downstream vision-language tasks, making them an obvious choice to solve Text-KVQA. Authors in (Yang et al., 2022; Khademi et al., 2023) prompt the LLMs with visual information via dense captions, object tags, object-level bounding box coordinates, and OCR tags. These methods rely heavily on the implicit knowledge learned by these LLMs. Further, KAT (Gui et al., 2022) improves upon such methods by augmenting external knowledge via retriever before prompting the LLM. However, it ignores the explicit visual information, which REVIVE (Lin et al., 2022) aims to fix. Although these methods show significant success, they have limitations such as hallucination and ignoring visual texts for reasoning. We aim to fill these gaps by proposing a novel solution for Text-KVQA.

Visual Entity Linking: Entity linking has traditionally been a well-established focus area within the NLP community (Jurafsky and Martin, 2009). In contrast, the problem of visual entity linking has only garnered attention in the last decade (Hu et al., 2023a; Sun et al., 2022; Shah et al., 2019). (Sun et al., 2022) have proposed a novel dataset and benchmark for visual named entity linking. (Shah et al., 2019) drew attention to the need for visual entity linking for addressing knowledge-based visual question answering. Open-domain Visual Entity Recognition has also been studied in the literature (Hu et al., 2023a; Caron et al., 2024; Xiao et al., 2024). However, most of these works have focused on linking entities such as persons, landmarks, and other named entities, while neglecting visual text such as business brand names and movie or book titles. In this work, we address this gap by proposing a principled solution for visual text entity linking and demonstrate its utility as a precursor to Text-KVQA.

3 Methodology

Problem Statement: Text-KVQA (Singh et al., 2019a) is a knowledge-intensive visual question-answering task that requires a system to read and interpret the visual text in an image and leverage it as a gateway to access and reason over external knowledge to answer the question. The external knowledge base \mathcal{K} consists of a set of n enti-



Figure 2: Challenges associated with Visual Text Entity Linking: (a) Visual text entity may appear as abbreviation instead of the entity name directly, e.g. “RBS” instead of “The Royal Bank of Scotland”, (b) Visual text with varying font and stylized orientation pose a challenge to the recognizer, (c) Example of homonyms where visual text *HP* may refer to ‘Hewlett Packard’ (left) or ‘Hindustan Petroleum’ (right).

ties $\mathcal{E} = \{E_1, E_2, \dots, E_n\}$ and their corresponding knowledge $\mathcal{K} = \{K_1, K_2, \dots, K_n\}$, where each K_i is a set of facts. For example, *Domino’s Pizza* is an entity whose associated knowledge facts, obtained in the form of triplets from Wikidata, are concatenated to form simple sentences such as “*Domino’s Pizza is a restaurant*”, “*Its headquarters are in Ann Arbor Charter Township*”, “*It belongs to the fast food industry*”, and so on. In this section, we describe our approach, whose overall architecture is illustrated in Figure 4. Our approach first links visual text entities using the proposed VisTEL module and retrieves relevant knowledge to the entity (Section 3.1), it then reasons over the image and the retrieved knowledge to answer the question (Section 3.2).

3.1 VisTEL: Visual Text Entity Linker

Entity linking is a well-studied task (Jurafsky and Martin, 2009), where given a sentence, the named entities need to be identified and linked with their corresponding entities in a knowledge base. In this work, we study an analogous task, where the input is no longer a sentence, but instead an image containing visual text entities and the task is to link them to a corresponding external knowledge base.

One plausible solution, as shown in (Singh et al., 2019a), is to extract the visual text in these images using visual text recognition engines and then lever-

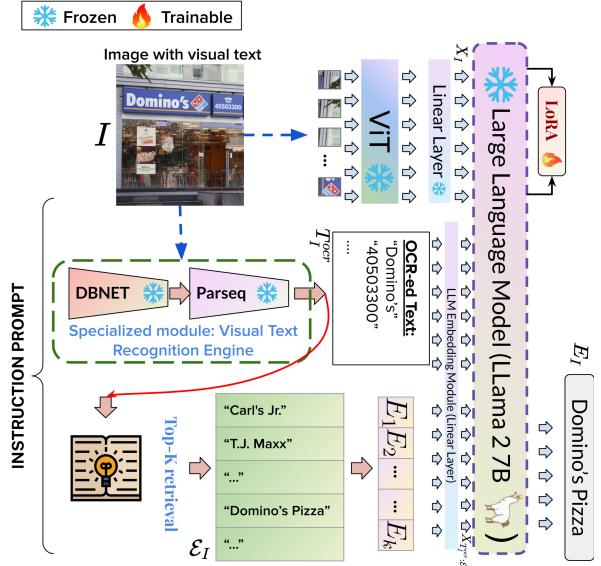


Figure 3: Illustration of VisTEL. We extract visual text from the given image using visual text recognition engine and, based on textual similarity, obtain k candidate entities from the knowledge base. We fit OCRed text and the candidate entities into an instruction prompt template and encode the image using a visual encoder and the text prompt using an LMM embedding module to obtain X_I and X_T , respectively. Once encoded, LMM generates the entity associated with the visual text in the image. Please refer to the Section 3.1.

age distance-based text similarity methods between the recognized text and the candidate entities for the entity linking task. However, such methods are highly sensitive to the following challenges: (i) Noisy or imperfect OCR may lead to wrong entity linking, and (ii) visual text might contain abbreviations instead of the entity names, e.g. “RBS” for the entity “The Royal Bank of Scotland”, (iii) The problem of homonymy, e.g. visual text *HP* may refer to ‘Hindustan Petroleum’ or ‘Hewlett Packard’. Furthermore, unlike entity linking which often benefits from larger textual contexts; visual text entity linking has limited textual context, e.g., surrounding visual texts, and often must infer correct entities based on visual context. Please refer to Figure 2 for a selection of challenges associated with visual text entity linking. The other plausible solution is to use large multimodal models (LMMs). By virtue of large-scale pretraining, they have strong abilities to reason and infer correct entities based on visual cues. However, we observe that feeding only the image without the surrounding OCRed text often results in hallucinations. To address these shortcomings, we propose Visual Text Entity Linker (VisTEL) that links the visual text present in an input image to its corresponding entity by jointly

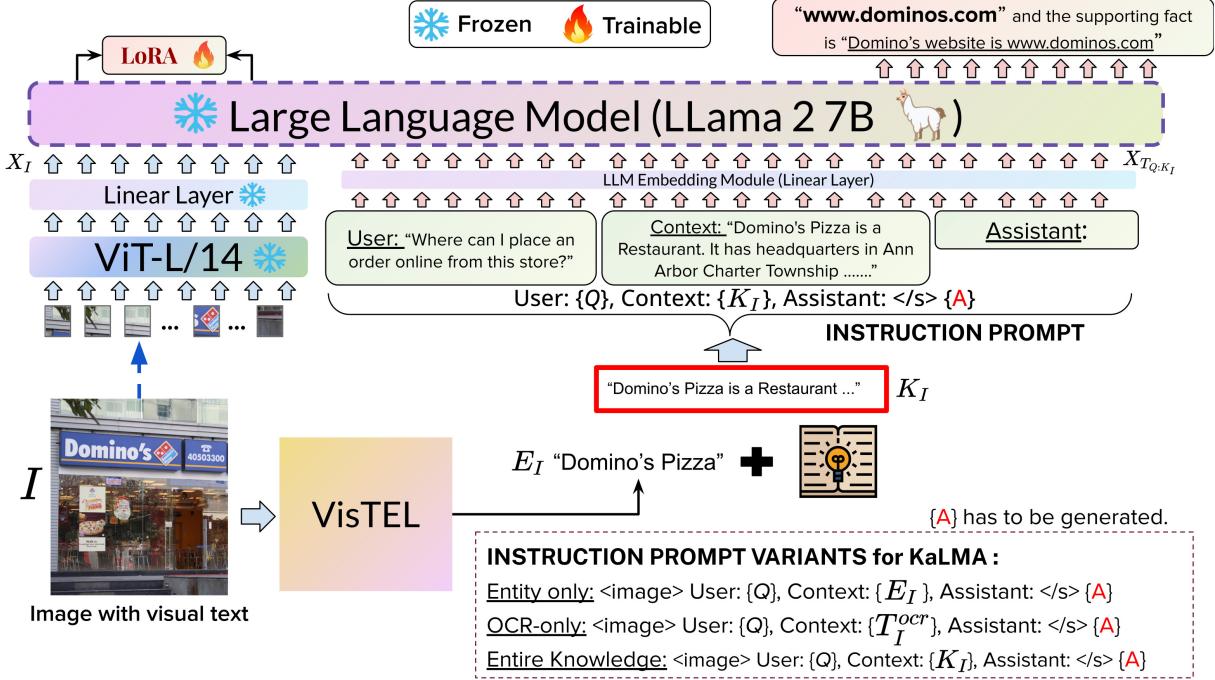


Figure 4: Overview of our proposed framework KaLMA. We first link the visual text in the image I to the entity E_I using VisTEL (Section 3.1) and its associated knowledge K_I is fetched. Then, we frame an instruction prompt with the question Q and the knowledge K_I , and encode it using the $lmm_{embedding}$ module f to obtain textual features $X_{T_{Q:K_I}}$. We encode the image I using a vision encoder to obtain visual features X_I . Then, we concatenate X_I and $X_{T_{Q:K_I}}$ and feed them to the LMM to generate an accurate answer A to the question Q . Instruction prompt templates used in our ablation study are shown in the bottom right box, where T_I^{ocr} is the visual text of the image I .

reasoning on textual context obtained using an explicit specialized visual text recognition engine and visual context obtained using a vision encoder of a large multimodal model. The architecture for VisTEL is illustrated in Figure 3.

Visual Text Recognition Engine: Given an image I , we extract text $T_I^{ocr} = \{t_1^{ocr}, t_2^{ocr}, \dots, t_r^{ocr}\}_I$ using specialized visual text detection and recognition methods. We, then find a set of k candidate entities \mathcal{E}_I based on the normalized edit-distance (NED) score between the entity name in the knowledge base with T_I^{ocr} . We use state-of-the-art text detection and text recognition approaches, namely DBNET (Liao et al., 2020) and ParSeq (Bautista and Atienza, 2022), respectively.

Vision encoder: We use the output of the last transformer layer of a pretrained CLIP visual encoder ViT-L/14 (Radford et al., 2021) as our patched image features $\tilde{X}_I \in \mathbb{R}^{p \times d_v}$, where p and d_v are the number of patches and encoding dimension of ViT, respectively. Further, these image features are projected to d_{lmm} dimension using a linear layer g to obtain the final sequence of image features $X_I \in \mathbb{R}^{p \times d_{lmm}}$, i.e., $X_I = g(\tilde{X}_I)$.

Large Multimodal Model: Once we obtain the OCR-ed text T_I^{ocr} and candidate entities \mathcal{E}_I , we

frame the following instruction prompt:

Instruction prompt template for VisTEL

```
<image>
USER: Given an image. The task is to link the visual text { $T_I^{ocr}$ } to one of the following entities: { $\mathcal{E}_I$ }
ASSISTANT: { $E_I$ }
```

Then, we feed the prompt to the embedding module h of the LMM to obtain text tokens $X_{T_I^{ocr}: \mathcal{E}_I} \in \mathbb{R}^{l \times d_{lmm}}$ i.e., $X_{T_I^{ocr}: \mathcal{E}_I} = h(prompt(T_I^{ocr} : \mathcal{E}_I))$, where l and d_{lmm} are the number of text tokens and input embedding dimension for the LMM, respectively. We, then concatenate image features X_I and text features $X_{T_I^{ocr}: \mathcal{E}_I}$, and feed it as an input to the large multimodal model. VisTEL auto-regressively predicts the probability of the next token E_{I_t} in the target entity E_I by attending to the input prompt tokens and the previously generated entity tokens $E_{I < t}$. We train VisTEL by optimizing the language modeling loss for generating the target entity conditioned on the inputs X_I and $X_{T_I^{ocr}: \mathcal{E}_I}$.

3.2 KaLMA: Knowledge-aware Large Multimodal Assistant

We present Knowledge-aware Large Multimodal Assistant (KaLMA) for addressing Text-KVQA.

The KalMA is an effective architecture that seamlessly integrates questions and images in the context of external knowledge in a trainable architecture to generate accurate answers.

We use visual features X_I from the vision encoder. Further, we concatenate question Q and the knowledge K_I via instruction prompt template (as shown in the Figure 4) and feed to the embedding module f of the LMM to obtain text tokens $X_{T_{Q:K_I}} \in \mathbb{R}^{m \times d_{lmm}}$ i.e., $X_{T_{Q:K_I}} = f(prompt(Q : K_I))$, where m is the number of text tokens. Then, we concatenate image features X_I , and text features $X_{T_{Q:K_I}}$ and feed to the large multimodal model to generate the accurate answer A . Further, to bring attribution ability, we model KalMA to generate the supporting fact S that contributed to the answer along with answer generation. From here onwards, we will refer answer and supporting fact together as A . KalMA predicts the probability of the next token A_{at} in the answer A_a in an auto-regressive manner. It does so by attending to the prompt inputs and the previously generated tokens $A_{a < t}$. We train by minimizing the generative language modeling loss $\mathcal{L}_{ans_gen}(\theta)$, which aims to generate the target tokens based on the inputs X_I and $X_{T_{Q:K_I}}$ (Eq. 1). Note that target tokens comprise both the answer and the supporting fact. During training, we leverage the ground truth entity and its corresponding knowledge K_I , while during inference, we obtain it using our VisTEL module. We reuse the weights of VisTEL to initialise KalMA.

$$\mathcal{L}_{ans_gen}(\theta) = - \left[\sum_{t=1}^{|A|} \log(P_\theta(A_{at} | A_{a < t}, X_I, X_{T_{Q:K_I}})) \right], \quad (1)$$

where θ are the trainable parameters, $A_{a < t}$ represents the answer tokens already generated before predicting the token A_{at} at the current time step t .

4 Experiments and Results

4.1 Dataset, Metrics and Comparisons

We conduct our experiments on Text-KVQA (Singh et al., 2019a) dataset². The questions in this dataset span across three splits, namely, scene, book, and movie containing natural scene images, book covers, and movie posters, respectively. These splits have (50K questions, 10K images, 500 entities), (1M questions, 207K images, 207K entities), (222K questions, 34K images, 34K entities),

| Method | Accuracy on Text-KVQA | | |
|--------------------------------------|--------------------------|--------------------------|--------------------------|
| | scene | book | movie |
| Traditional VQA Baselines | | | |
| BiLSTM | 17.0 | 12.4 | 11.3 |
| BoW+CNN | 11.5 | 8.7 | 7.0 |
| BLSTM+CNN (Antol et al., 2015) | 19.8 | 17.3 | 15.7 |
| HiCoAttenVQA (Lu et al., 2016) | 22.2 | 20.4 | 18.4 |
| BAN (Kim et al., 2018) | 23.5 | 22.3 | 20.3 |
| Pre-LMM Approaches | | | |
| GPT-2 (Radford et al., 2019) | 22.8 | 22.3 | 31.8 |
| GPT-2 (w/ Visual Context) | 25.4 | 43.2 | 38.5 |
| ViLT (Kim et al., 2021) | 38.2 | 31.1 | 40.1 |
| VLBart (Cho et al., 2021) | 35.1 | 38.6 | 41.5 |
| Previous SOTA | | | |
| Memory Network (Weston et al., 2015) | 49.0 | 57.2 | 42.0 |
| Singh et al. (Singh et al., 2019a) | 54.5 | 62.7 | 45.2 |
| LLM-based Approaches | | | |
| mPlug-Owl (Ye et al., 2023) | 21.3 | 26.7 | 8.2 |
| LLaVA-1.5 (Liu et al., 2024) | 39.2 | 37.0 | 46.1 |
| MiniGPT4v2 (Zhu et al., 2023) | 48.2 | 47.7 | 47.6 |
| InstructBLIP (Dai et al., 2024) | 31.5 | 30.3 | 29.9 |
| Ours (KalMA) | | | |
| w/ NED retrieval | 54.9 | 63.4 | 70.8 |
| w/ VisTEL | 72.7 (\uparrow 18.2%) | 82.3 (\uparrow 19.6%) | 77.4 (\uparrow 32.2%) |
| Oracle | 99.3 | 92.8 | 99.4 |

Table 1: **Results on Text-KVQA:** Various methods on the three data categories of Text-KVQA dataset, namely, scene, book and movie.

respectively. Further, each of these splits comes with its own knowledge base, namely *KB-business* containing knowledge facts about business brand entities harvested from Wikidata, *KB-book* containing knowledge facts about books harvested from a book catalog, and *KB-movie* containing knowledge facts about movies harvested from IMDB, respectively. For each split, we follow the similar train-test division as (Singh et al., 2019a) where entities in train and test sets are disjoint. We evaluate the methods using an accuracy metric.

Along with traditional VQA baselines, we compare the question answering performance of our proposed approach KalMA with methods from the following three major categories: (i) **Pre-LMM Approaches:** here, we choose classical transformer-based baselines, namely, GPT-2 (Radford et al., 2019) (text-only), GPT-2 (with BLIP-2 (Li et al., 2023)-extracted captions as visual context), ViLT (Kim et al., 2021) and VLBart (Cho et al., 2021). For an encoder-only model like ViLT, we treat Text-KVQA as a classification-style visual question answering where the task is to predict the answer from a set of all possible answers. (ii) **LMM-based Approaches:** restricting ourselves to open-source models, we choose four popular LMMS, namely, mPlug-Owl (Ye et al., 2023), MiniGPT4v2 (Zhu et al., 2023), LLaVA-1.5 (7B) (Liu et al., 2024) and InstructBLIP (Dai et al., 2024) for comparison. Prompts used and other fine-tuning details for these LMMS are discussed in the Appendix. (iii) **SOTA approaches:** we also compare against memory network (Weston et al., 2015)

²Available at: <https://textkvqa.github.io>

| | | | | |
|--------------|--|--|---|--|
| | (a) | (b) | (c) | (d) |
| Input Image | | | | |
| Question | Which retail store is this? | Which year this was founded? | Who is the director of this movie? | What is the title of this book? |
| Ground Truth | T.J. Maxx | 1927 | Danny Boyle | Meteorites |
| LLaVA-1.5 | Target Corporation X | 1976 X | James Franco X | The Geology of Meteorites X |
| MiniGPT4v2 | Target X | 1995 X | James Cameron X | Metamorphic Rocks X |
| InstructBLIP | Retail Store X | 1991 X | Tim Story X | Meteors X |
| mPLUG-Owl | 99p Stores X | 1971 X | 1 X | The history and mystery of the most important natural phenomenon of the last 1000 years meteorites asteroids and comets X |
| KaLMA (Ours) | T.J. Maxx ✓ [AND] the supporting fact is 'This retail store is T.J. Maxx' | 1927 ✓ [AND] the supporting fact is 'Z-Eleven was established in 1927.' | Danny Boyle ✓ [AND] the supporting fact is '127 Hours movie is directed by Danny Boyle.' | Meteorites ✓ [AND] the supporting fact is "This is 'Meteorites' book." |
| | (e) | (f) | (g) | (h) |
| Input Image | | | | |
| Question | Which retail store is this? | Which bank is this? | Which retail store is this? | Where is it headquartered? |
| Ground Truth | B&Q | Jyske Bank | Franprix | Dallas |
| LLaVA-1.5 | Tesco PLC X | Svenska handelsbanken X | Carrefour X | New York X |
| MiniGPT4v2 | B&Q ✓ | Bnp paribas X | Carrefour X | Chicago X |
| InstructBLIP | Five Below X | Bank of sweden X | Retail store X | Thermal City X |
| mPLUG-Owl | 99p Stores X | Orsted X | 7-eleven X | 100 west Madison Street, Chicago X |
| KaLMA (Ours) | B&Q ✓ [AND] the supporting fact is 'This retail store is B&Q' | Jyske Bank ✓ [AND] the supporting fact is 'This bank is Jyske Bank' | Franprix ✓ [AND] the supporting fact is 'This retail store is franprix' | Dallas ✓ [AND] the supporting fact is "Tuesday Morning has headquarters in Dallas." |

Figure 5: A selection of our results as compared to implicit knowledge-based LMM approaches. Please refer Qualitative Results in Section 4.3 for observations. More results in Appendix C.

and graph neural network-based approach (Singh et al., 2019b) which are the current state of the art. In addition to these comparisons, we compare the visual text entity linking performance of our proposed VisTEL against recent multimodal retrievers from UniIR (Wei et al., 2024), specifically CLIP-SF and BLIP-SF, where we use image and visual text to retrieve entities from the knowledge base.

4.2 Implementation Details

We implemented our method using PyTorch and the Huggingface Transformers library (Wolf et al., 2020). We used LLaVA-1.5 as our foundation model for both VisTEL and KaLMA models. Note that, LLaVA-1.5 is trained on CC3M (Sharma et al., 2018) and MS-COCO (Lin et al., 2014). We have carefully examined these datasets for duplicates and found no overlap with the evaluation set of Text-KVQA. Further, DBNET (Liao et al., 2020) and PARSEQ (Bautista and Atienza, 2022) are used as visual-text detection and visual-text recognition

modules in the visual text recognition engine, respectively. We fine-tuned VisTEL with LoRA for 10 epochs with a learning rate of 1e-5 with a batch size of 128. Similarly, we fine-tuned KaLMA with LoRA for 6 epochs with a learning rate of 2e-5 with a batch size of 64. LoRA details are as follows: rank: 16, alpha: 32, dropout: 0.05, for both the models. Our experiments are conducted on a machine with three A6000 GPUs (48 GB each). We make our implementation publicly available at our project website³.

4.3 Results and Discussion

Results on Text-KVQA: We quantitatively evaluate our proposed framework KalMA on Text-KVQA and compare against relevant methods in Table 1. We report accuracy averaged over the entire test set for all the three splits of Text-KVQA. It is no surprise that traditional VQA baselines perform poorly as they do not have the ability to read and

³<https://vl2g.github.io/projects/LMM4Text-KVQA/>



Figure 6: Comparison of visual text entity linking results. The VisTEL infers the correct entity based on visual context as well as textual context in the form of surrounding text in the image. Please refer to Qualitative Results in Section 4.3 for more details.

reason over visual text. Pre-LMM language models (GPT-2) and vision-language models (GPT-2 w/ visual context, ViLT, VisualBert) along with LMM baselines (mPlug-Owl, LLaVA-1.5, MiniGPT4v2, InstructBLIP) outperform traditional methods, but fail to outperform knowledge-aware methods including the state-of-the-art method (Singh et al., 2019b). We observe that on knowledge-intensive tasks like Text-KVQA, the OCR-free capabilities acquired by LMMS are due to heavily correlated hallucinations of visual objects, thereby fall short to our proposed approach by a significant margin. Our proposed framework seamlessly integrates knowledge associated with visual text entity (extracted using our proposed VisTEL) and significantly enhances the performance on Text-KVQA. To be specific, we advance the state-of-the-art by 18.2%, 19.6%, and 32.2% on scene, book, and movie splits of Text-KVQA on an absolute scale. This superiority of our approach demonstrates its efficacy in knowledge-aware text-based visual question answering.

Visual Text Entity Linking Results: We report them in Table 2. Here, we observe that the proposed VisTEL clearly outperforms both (i) Text-only retrievers, such as a direct match or normalized edit distance-based match of OCRed text and entity name, and (ii) Multimodal retrievers, CLIP-SF and BLIP-SF from UniIR (Wei et al., 2024). By

| Method | Visual Context | Textual Context | scene | book | movie |
|------------------------------|----------------|-----------------|-------------|-------------|-------------|
| Text-only | | | | | |
| Direct match | x | ✓ | 54.8 | 63.6 | 58.1 |
| NED | x | ✓ | 57.1 | 66.5 | 60.1 |
| Multimodal retrievers | | | | | |
| UniIR (CLIP-SF) | ✓ | ✓ | 64.5 | 78.8 | 45.2 |
| UniIR (BLIP-SF) | ✓ | ✓ | 60.6 | 78.5 | 50.1 |
| Ours | | | | | |
| VisTEL | ✓ | ✗ | 73.2 | 76.9 | 66.6 |
| VisTEL | x | ✓ | 31.5 | 9.8 | 11.6 |
| VisTEL | ✓ | ✓ | 76.5 | 80.6 | 71.6 |

Table 2: Visual Text Entity Linking Results. We report Recall@1. Text-only retrievers: direct match and normalized edit distance-based methods and Multimodal retrievers: CLIP-SF and BLIP-SF from UniIR (Wei et al., 2024) fall short. On the contrary, the proposed VisTEL, which leverages both visual and textual context (surrounding OCRed text) in an LMM framework, shows impressive visual text entity linking performance over both text-only as well as multimodal retrievers.

virtue of LMM and joint reasoning of visual and textual (OCR) context for linking visual text, VisTEL yields reasonably advanced performance. Nevertheless, there is still scope of improvement which we believe can be achieved by further improving visual text recognition, and performing detailed visual reasoning such as logo recognition. We leave these extensions as future work.

Qualitative Results: We show a selection of results for text-based knowledge-aware visual question answering and visual text entity linking in Figure 5 and Figure 6, respectively.

In Figure 5, LMM models exhibit hallucination over visually apparent objects. In (a), all LMMs incorrectly identify *T.J. Maxx* as popular retail stores *Target* and *99p Stores*. In (b), they provide a random year. In (c), these models are confused over the keyword *James*, mixing up the director and actor names on the poster. In (d), LMMs hallucinate and suggest non-existent book titles. Similar hallucinations can be seen in the other examples (e-h). Our proposed method owing to visual-text entity linking capabilities and reasoning over explicit knowledge, provides accurate answers.

In Figure 6, we observe that our proposed model accurately links visual text in the images to the correct entity despite noisy OCR in (a), abbreviations in (b), and ambiguous visual text in (c).

4.3.1 Ablations and Analysis

We conduct the following ablations and analysis of the proposed work:

(i) **What is the need for VisTEL?**: To study the performance of our model in the absence of the proposed VisTEL module, we replace it with traditional edit-distance-based entity linking where

| Model | Visual text EL | Knowledge | scene | book | movie |
|------------------------------------|----------------|------------------|-------------|-------------|-------------|
| KaLMA | x | x | 39.2 | 37.0 | 46.1 |
| | x | OCR only | 52.2 | 49.8 | 51.7 |
| | ✓ | Entity name only | 53.2 | 59.1 | 59.2 |
| | ✓(w/o VisTEL) | Knowledge facts | 54.9 | 63.4 | 70.8 |
| -- MinIGPT4v2 (best LMM method) | ✓(w/ VisTEL) | Knowledge facts | 72.7 | 82.3 | 77.4 |
| | x | x | 48.2 | 47.7 | 47.6 |

Table 3: Ablations for showing the importance of visual text entity linking, explicit knowledge facts and VisTEL. Also, note that the first-row result corresponds to LLaVA-1.5 result from Table 1, as KaLMA without VisTEL and knowledge is equivalent to LLaVA-1.5. Please refer to Section 4.3.1 for more details.

| Detection | Recognition | scene | book | movie |
|-----------|-------------|-------------|-------------|-------------|
| EAST | CRNN | 67.2 | 81.3 | 66.4 |
| CRAFT | CRNN | 67.7 | 81.9 | 75.1 |
| DBNet | ParSeq | 72.7 | 82.3 | 77.4 |

Table 4: Effect of Different Text Detection and Recognition Approaches in our approach.

entities are sorted based on the normalized edit-distance between extracted OCRs and the entity name. The results of this ablation, as shown in Table 3 further support our claim that the superior visual text entity linking capabilities of the proposed VisTEL, enhances the downstream performance of KaLMA.

(ii) What is the need for visual text entity linking and explicit knowledge in Text-KVQA?: We show these ablation results in Table 3. We, **first**, skip visual entity linking in the KaLMA, and feed only the extracted OCRed text to KaLMA. The drop in performance shows the utility of visual text entity linking. **Second**, we perform visual entity linking, but, we feed the visual text linked entity name from the VisTEL as input to KaLMA. Our observations indicate that although entity names give some hints about the associated knowledge and reduce hallucination to some extent, it is not as useful as using explicit knowledge in our full model.

(iii) How much does choice of visual text recognition engine matter?: In this ablation, we replace DBNET (Liao et al., 2020) and ParSeq (Bautista and Atienza, 2022) used in KaLMA with CRAFT (Baek et al., 2019), EAST (Zhou et al., 2017) and CRNN (Shi et al., 2016), and report the results of KaLMA on Text-KVQA in Table 4. Although effective visual text recognition is critical to the performance, our model that jointly reasons on visual and textual context, performs reasonably well even with sub-par visual text recognition.

(iv) Attribution ability of KaLMA: To study the impact of support fact generation (SFG) along with the answer generation on the performance of KaLMA, we train KaLMA without support fact

| SFG | scene | book | movie |
|-----|-------------|-------------|-------------|
| ✓ | 72.7 | 82.3 | 77.4 |
| x | 71.4 | 83.5 | 76.9 |

Table 5: Performance of KaLMA w/ and w/o supporting fact generation (SFG) on Text-KVQA.

generation, and report the results in Table 5. We observe that KaLMA’s performance drops slightly, further supporting our claim that support fact generation elicits chain-of-thought reasoning, thereby improving the performance of answer generation along with adding attribution abilities to the model.

(iv) Cost analysis: We provide a comparison of KaLMA with a traditional non-LLM-based approach (ViLT). Our approach takes on average 5.6s per sample, which includes 4s for visual text recognition, 0.8s for entity linking using VisTEL and 0.8s for VQA using KaLMA as compared to ViLT which takes on average 0.2s per sample during inference. The training time (finetuning) of both these models are 36 and 8 hrs, respectively. Furthermore, the trainable parameters for both these models are 20M (Total size: 14B) and 114M (Total size: 114M), respectively. We achieved speed-up in our LMM components through parameter-efficient fine-tuning (LoRA) with 16-bit precision and 8-bit quantization during inference. As anticipated, traditional models have a notable advantage in terms of computational efficiency compared to our LMM-based approach. Nonetheless, we substantially surpass them in Text-KVQA accuracy.

5 Conclusion

We have revisited the Text-KVQA and significantly advanced state of the art on this task. Our findings suggest that visual text entity linking, combined with seamless reasoning using both visual and textual cues, as well as explicit external knowledge via LMM, is key to our success. We performed extensive ablation studies and analyses to support our claims. The future scope of this work is to expand the dataset with more visual-intensive queries and address Text-KVQA for multilingual societies.

6 Limitations

We observe the following limitations in our work: (i) Existing visual text recognition pipelines suffer on low-resolution images where it is challenging to extract visual text, which further impacts the performance of our VisTEL (ii) In the dataset we use, it was assumed that each image contains only

one visual text entity which may not be always true in a real-world scenario. (iii) Current state-of-the-art visual text recognition engines are not effective enough over multi-lingual text in the wild; Hence, in this work, we further assume the visual-text is English which again might not hold in a realistic setting. (iv) The temporal nature of knowledge, such as the entity “Statoil” being renamed “Equinor” over time, is not handled by our current models. We leave addressing these limitations as a future work of this paper.

7 Ethical Considerations and Broader Impact

This work is based on the publicly available Text KVQA dataset, which predominantly contains English visual text, and the associated knowledge base, questions, and answer pairs are also in English. The dataset may have some geographic bias that went undetected in this work, a common issue with many public computer vision and NLP benchmarks. Additionally, our work uses large multimodal models (LMMs), which can inherit and potentially amplify biases from the large-scale pre-training data used.

We are mindful of the environmental impact of using LMMs due to their heavy computational requirements. To mitigate this, we judiciously used LMMs by reusing pre-existing checkpoints wherever appropriate.

We open-source our implementation to facilitate reproduction and further study. Nevertheless, a more rigorous inspection is indeed required before deploying the proposed model in real-world applications to ensure ethical considerations are comprehensively addressed.

Broader Impact: The proposed work has the following broader impact: (i) The ability to link visual text entities to knowledge bases and leverage this linked knowledge for answering questions can improve the accuracy and relevance of information retrieval systems. Although not studied in this work, this may be particularly valuable in content recommendation systems and search engines. (ii) This research contributes to advancing the capabilities of AI systems to understand and interact with multimodal information (text and images), which can benefit applications in fields such as virtual assistants, content understanding, and automated decision-making. (iii) Methodologically, contributions such as VisTEL provide new frameworks and

techniques for visual text entity linking, which can inspire further innovations in Visual NLP.

Acknowledgements

This work was partly supported by the IIT Jodhpur Seed Research Grant and National Language Translation Mission (NLTM): Bhashini project by the MeitY, Government of India. Abhirama Subramanyam Penamakuri was supported by the PMRF fellowship, MoE, Government of India.

References

- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *ICCV*.
- Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019. Character region awareness for text detection. In *CVPR*.
- Darwin Bautista and Rowel Atienza. 2022. Scene text recognition with permuted autoregressive sequence models. In *ECCV*.
- Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. 2019a. Icdar 2019 competition on scene text visual question answering. In *ICDAR*.
- Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019b. Scene text visual question answering. In *ICCV*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *NeurIPS*.
- Mathilde Caron, Ahmet Iscen, Alireza Fathi, and Cordelia Schmid. 2024. A generative approach for wikipedia-scale visual entity recognition. In *CVPR*.
- Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? In *EMNLP*.
- Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *ICML*.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructclip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.
- François Gardères, Maryam Ziaeefard, Baptiste Abeiloos, and Freddy Lecue. 2020. ConceptBert: Concept-aware representation for visual question answering. In *EMNLP*.
- 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 *CVPR*.
- Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander G Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. Kat: A knowledge augmented transformer for vision-and-language. In *NAACL-HLT*.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *ICML*.
- Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. 2023a. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *ICCV*.
- Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. 2023b. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *CVPR*.
- Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., USA.
- Mahmoud Khademi, Ziyi Yang, Felipe Vieira Frujeri, and Chenguang Zhu. 2023. Mm-reasoner: A multi-modal knowledge-aware framework for knowledge-based visual question answering. In *EMNLP*.
- Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *NeurIPS*.
- Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*.
- 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. In *ICML*.
- Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. 2020. Real-time scene text detection with differentiable binarization. In *AAAI*.
- 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 *ECCV*.
- Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. 2022. Revive: Regional visual representation matters in knowledge-based visual question answering. In *NeurIPS*.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *CVPR*.
- Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *NeurIPS*.
- Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based VQA. In *CVPR*.
- Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *CVPR*.
- Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. DocVQA: A dataset for VQA on document images. In *WACV*.
- Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. OCR-VQA: Visual question answering by reading text in images. In *ICDAR*.
- Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. In *NeurIPS*.
- Medhini Narasimhan and Alexander G Schwing. 2018. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In *ECCV*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2024. The refinedweb dataset for falcon LLM: Outperforming curated corpora with web data only. In *NeurIPS*.

- 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 *ICML*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. KVQA: Knowledge-aware visual question answering. In *AAAI*.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.
- Baoguang Shi, Xiang Bai, and Cong Yao. 2016. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE TPAMI*, 39(11):2298–2304.
- Ajeet Kumar Singh, Anand Mishra, Shashank Shekhar, and Anirban Chakraborty. 2019a. From strings to things: Knowledge-enabled VQA model that can read and reason. In *ICCV*.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019b. Towards VQA models that can read. In *CVPR*.
- Wen Sun, Yixing Fan, Jiafeng Guo, Ruqing Zhang, and Xueqi Cheng. 2022. Visual named entity linking: A new dataset and A baseline. In *EMNLP (Findings)*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwala Bhargava, Shruti Bhosale, et al. 2023. 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. In *NeurIPS*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.
- Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017a. FVQA: Fact-based visual question answering. *TPAMI*, 40(10):2413–2427.
- Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. 2017b. Explicit knowledge-based reasoning for visual question answering. In *IJCAI*.
- Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhui Chen. 2024. Uniir: Training and benchmarking universal multimodal information retrievers. In *ECCV*.
- Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. Memory networks. In *ICLR*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrette Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP: System Demonstrations*.
- BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Lucioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.
- Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *CVPR*.
- Zilin Xiao, Ming Gong, Paola Cascante-Bonilla, Xingyao Zhang, Jie Wu, and Vicente Ordonez. 2024. Grounding language models for visual entity recognition. *arXiv preprint arXiv:2402.18695*.
- 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 *AAAI*.
- Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Deegan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.
- Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander Smola, and Le Song. 2018. Variational reasoning for question answering with knowledge graph. In *AAAI*.
- Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. East: an efficient and accurate scene text detector. In *CVPR*.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Appendix

A Question Categorisation

We show the visual question-answering results over concretized sub-categories under each of the scenes, book and movie split in Table 6. We observe that our proposed model shows remarkable performance across diverse question categories, particularly in the challenging categories such as date, people, and open-ended question categories.

B Finetuning details of LMMs

In this section, we explain the hyperparameters and prompts used to finetune the LMMs. Note that we conduct all our experiments on a machine with 3 48GB A6000 GPUs. For mPlug-Owl and MiniGPT4v2, we have used hyperparameters as per the original papers.

mPlug-Owl: We finetuned mPlug-Owl with LoRA for 6 epochs with a learning rate of 2e-5 with a batch size of 256. LoRA details: rank: 8, alpha: 32, dropout: 0.05.

Instruction prompt template for mPlug-Owl

The following is a conversation between a curious human and an AI assistant. The assistant gives accurate and crisp answers to the user’s questions.

Human: <image>

Human: {Q}

AI: {A}.

MiniGPTv4v2: We finetuned MiniGPTV4v2 with LoRA for 6 epochs with a learning rate of 3e-5 with a batch size of 128. LoRA details: rank: 16, alpha: 64, dropout: 0.05.

Instruction prompt template for MiniGPT4v2

<image>

{vqa} Based on the image, respond to this question with a short answer: {Q}, ASSISTANT: {A}

InstructBLIP: We finetuned InstructBLIP for 3 epochs with a learning rate of 1e-5 with a batch size of 128.

Instruction prompt template for Instruct-BLIP

<image>

USER: {Q}. ASSISTANT: {A}

LLaVA-1.5: We finetune LLaVA with LORA for 6 epochs with a learning rate of 5e-5 with a batch size

of 64. LoRA details: rank: 16, alpha: 32, dropout: 0.05.

Instruction prompt template for LLaVA-1.5

<image>

USER: {Q}. ASSISTANT: {A}

C More Results

More qualitative results on movie and book splits of Text-KVQA are shown in Figure 7 and Figure 8, respectively.

| Method | Text-KVQA (scene) | | | | | Text-KVQA (book) | | | | | Text-KVQA (movie) | | | | | |
|---------------------------|-------------------|-------------|-------------|-------------|-------------|------------------|-------------|-------------|-------------|-------------|-------------------|-------------|-------------|-------------|-------------|-------------|
| | B | D | P | L | OE | B | D | P | G | OE | B | D | P | G | L | OE |
| Pre-LLM Methods | | | | | | | | | | | | | | | | |
| GPT-2 | 54.8 | 0.2 | 0.0 | 13.7 | 15.4 | 54.5 | 43.8 | 0.1 | 4.3 | 0.6 | 74.5 | 2.1 | 0.0 | 15.2 | 63.7 | 0.0 |
| GPT-2 (w/ Visual Context) | 57.1 | 0.3 | 0.0 | 16.1 | 17.0 | 80.1 | 63.8 | 5.2 | 45.1 | 7.5 | 75.4 | 3.2 | 0.0 | 24.3 | 66.8 | 29.3 |
| ViLT | 75.9 | 0.0 | 0.0 | 33.9 | 28.7 | 68 | 63.3 | 0 | 21.3 | 0.9 | 85 | 4.4 | 0.2 | 42.1 | 76.7 | 0.0 |
| VLBart | 78.9 | 0.2 | 0.0 | 18.8 | 27.4 | 79.2 | 62.0 | 1.7 | 34.9 | 0.9 | 85.4 | 6.3 | 0.0 | 43.7 | 76.7 | 0.0 |
| LLM Methods | | | | | | | | | | | | | | | | |
| mPlug-Owl | 22 | 8.9 | 0.0 | 45 | 9.8 | 19.5 | 69.7 | 38.7 | 43.8 | 12 | 7.8 | 17.5 | 0.7 | 9.7 | 6.2 | 5.5 |
| LLaVA-1.5 | 81.1 | 0.0 | 2.0 | 38.7 | 23.4 | 79 | 70.6 | 19.3 | 57.3 | 2.7 | 84.8 | 13.5 | 0.3 | 1.6 | 72.7 | 9.9 |
| MiniGPT4v2 | 81.7 | 2.7 | 1.3 | 49.9 | 41.7 | 80.1 | 71.9 | 18.2 | 54.2 | 6.6 | 79.9 | 13.6 | 1.2 | 53.7 | 78.4 | 30.4 |
| InstructBLIP | 50.0 | 0.1 | 6.6 | 29.7 | 32.8 | 49.8 | 70.3 | 22 | 15.2 | 12.8 | 50.0 | 6.6 | 0.3 | 1.4 | 76.5 | 39.5 |
| Ours | | | | | | | | | | | | | | | | |
| KaLMA | 77.2 | 69.0 | 76.8 | 67.8 | 69.9 | 88.5 | 72.9 | 80.0 | 80.2 | 79.6 | 84.2 | 69.6 | 74.8 | 70.6 | 91.5 | 69.1 |
| KaLMA (Oracle) | 83.9 | 95.8 | 95.4 | 91.9 | 91.8 | 98.0 | 96.4 | 98.2 | 99.9 | 98.2 | 99.9 | 99.8 | 95.9 | 100.0 | 100.0 | 99.7 |

Table 6: QA accuracy performance breakdown for various methods by question categories on TEXT-KVQA. Categories are **B**: binary, **D**: date, **P**: people, **L**: location, **G**: genre and **OE**: open-ended.

| | | | |
|---|---|---|--|
|  |  |  |  |
| Input Image | | | |
| Question | Who is the director of the movie? | In which year this movie was released? | Is this a biography? |
| Ground Truth | Michele Lupo | 2015 | No |
| LLaVA-1.5 | John Huston X | 1975 X | Yes X |
| MiniGPT4v2 | John Cleese X | 2014 X | Yes X |
| InstructBLIP | Shin'ichirō Sawazaki X | 2013 X | No ✓ |
| mPLUG-Owl | Nigoi X | 2010 X | Yes X |
| KaLMA (Ours) | Michele Lupo ✓ [AND] the supporting fact is 'Sette volte sette is directed by michele lupo.' | 2015 ✓ [AND] the supporting fact is 'Welcome to Leith was released in 2015.' | No ✓ [AND] the supporting fact is 'The sound and the shadow is a comedy movie.' |
| (a) | (b) | (c) | (d) |

Figure 7: A few more selection of our results as compared to implicit knowledge-based LMM approaches on the movie subset of Text-KVQA.

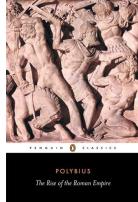
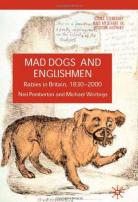
| | | | |
|---|--|---|--|
|  |  |  |  |
| Input Image | | | |
| Question | Who wrote this book? | Is this a games realldl book? | Is this a pharmaceutical book? |
| Ground Truth | Polybius | No | Yes |
| LLaVA-1.5 | Anonymous X | Yes X | No X |
| MiniGPT4v2 | Pylos X | Yes X | No X |
| InstructBLIP | Sophus Clausius X | No ✓ | No X |
| mPLUG-Owl | Strabo X | Yes X | Yes ✓ |
| KaLMA (Ours) | Polybius ✓ [AND] the supporting fact is 'The rise of the roman empire (penguin classics) is written by polybius.' | No ✓ [AND] the supporting fact is 'The great cholesterol con: the truth about what really causes heart disease and how to avoid it book's genre is health, fitness and dieting.' | No ✓ [AND] the supporting fact is 'The sound and the shadow is a comedy movie.' |
| (a) | (b) | (c) | (d) |

Figure 8: A few more selection of our results as compared to implicit knowledge-based LMM approaches on the book subset of Text-KVQA.