Supplementary Materials: AdQuestA: Knowledge-Guided Visual Question Answer Framework For Advertisements

Neha Choudhary, Poonam Goyal, Devashish Siwatch, Atharva Chandak, Harsh Mahajan
APPCAIR AI Research Centre, Disruptive Technologies Lab, BITS Pilani, India
{p20190409, Poonam, f20200113, f20190062, f20190036}@pilani.bits-pilani.ac.in

Varun Khurana, Yaman Kumar
Media and Data Science Research Lab, Adobe, India
{varunkhurana, ykumar}@adobe.com

1. Overview

We provide the supplementary material in four section: (a) Details of the dataset in Appendix A (b) Example prompts which are used for MIKG module and finetuning module in Appendix B. (c) Details of competitive models in Appendix C. (d) Qualititive analysis figure in Appendix D. (e) Visualization of success and failure cases in Appendix E.

We have created $94,811$ QA pairs for $17,118$ ad images sourced from Facebook, twitter and dataset given in Hussain et al. [2]. We annotate the dataset for total 15 unique questions and pose 5-9 questions per image. List of the questions is given in Table 1 and the distribution of these questions in our dataset is shown in Figure 1.

We have two type of questions: 1) Categorised: answers to these questions belong to predefined categories and 2) Non-categorised: answers to these questions are either descriptive or one/multi words but open-ended in nature. For example, Q11 (in Table 1): What is the sentiment of this ad? Answer for this question can belong to one of the categories given in the same table such as active, afraid, alarmed, etc. Similarly, for questions Q13 and Q15, answers are also categorised and categories are taken from the Hussain et al. [2] dataset. The questions, Q12: This ad is trustworthy or not and why? and Q14: What is the persuasion of this ad?, are taken from Kumar et al. [3] and Zeng et al. [7], respectively. They have also provided answer categories. Our annotators have used these predefined categories to annotate answers to respective questions.

1.1. Appendix A

The answers for the questions Q1-Q10 are open-ended descriptive or single/multi words. Few answer samples are given for each question in Table 1. We have taken Q6 and Q7 from Hussain et al. [2] dataset. Other questions in this

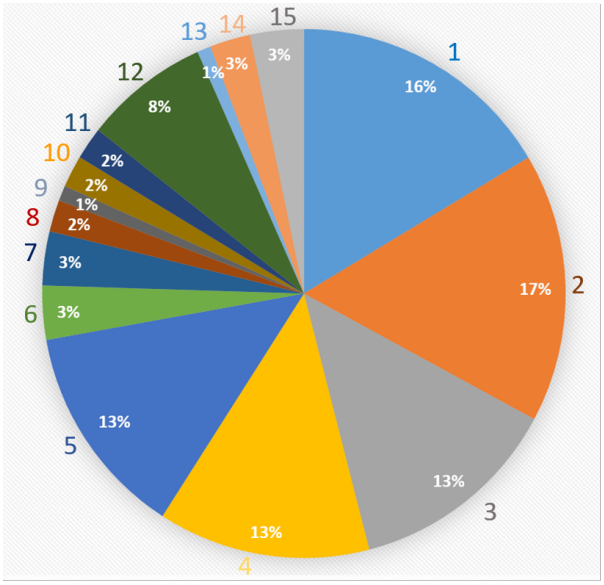

Figure 1. Distribution of questions in ADVQA dataset; $x\%$ denotes the x percentage of QA pairs having that question. Numbers 1-15 correspond to the questions listed in Table 1

category are created by us. Our annotators have annotated answers for these questions using our instructions and sample answers. The sample context for the given ad image in Figure2 (these passages are ranked by QACR module of AdQuestA based on the question posed). The question posed in this example is "This advertisement is related to?". It is evident from the ad image that achieving ground truth answer "Casualwear, hosiery and Socks, Innerwear" is difficult only from visuals. It can be seen from the highlighted portion of the ad context that the answer is available in some form.

Table 1. List of questions in ADVQA with sample answers/answer categories

S.No.	Question	Example
		Non-Categorical
1	What is the company name?	Nike; Amazon; Pepsi; Starbucks; darden restaurants; Nivea .....etc.
2	What is the title of the company?	Cardinal Health, Frontier Communications, Dollar Tree .....etc.
3	What is the description of the advertisement?	Clinical laboratory company; Rocket Companies is a Detroit-based fintech company consisting of mortgage, real estate and financial service businesses.; Waste disposal corporation .....etc.
4	This advertisement is related to?	Retail Trade Sector; Automotive industry; Tobacco; Healthcare; Textile; Electric Utilities; Video games.....etc.
5	What type of products does the company sell?	Clothing; Health plans; Transport; Food and Snacks; Hair .....etc.
6	What should i do according to this ad?	I should buy heinz ketchup, I should buy grey poupon mustard, I should buy an hp printer..... etc.
7	Why should i follow this ad?	it's a cool car,it's fast,they have been in business for 30 years it's what tastes right,it's open late in the nights.....etc.
8	In which year this company is established?d	1858; 1974; 1888; 1990 .....etc.
9	Who is the target audience?	Race Enthusiasts; Business Women; Movie Lovers; Middle-aged; Patients, Education Seekers.....etc.
10	What is the slogan of the company?	Thinking Beyond Price.; Solutions for a Sustainable World; Life's Better When We're Connected; The Future Is Fusion...etc.
		Categorical
11	What is the sentiment of this ad?	Active; Afraid; Alarmed; Alert ; Amazed ; Amused; Angry; Calm; Cheerful; Confident; Conscious; Creative; Disturbed; Eager; Educated; Emotional; Empathetic; Fashionable; Feminine; Grateful; Inspired; Jealous; Loving; Manly; Persuaded; Pessimistic; Proud; Sad; Thrifty; Youthful
12	This ad is trustworthy or not and why?	Bad ads: Boring; Irrelevant,Cheap; Ugly; Badly Designed, Clickbait;Deceptive; Untrustworthy; Don't Like the Product ; Offensive, Uncomfortable,Distasteful; Politicized; Pushy, Manipulative; Unclear; Good ads: Entertaining,Engaging; Good Style and/or Design; Interested in the Product or Topic; Simple, Straightforward; Trustworthy, Genuine; Useful, Interesting, Informative
13	What is the atypical category?	Texture Replacement 1; Texture Replacement 2; Object with Missing Part;Combination of Parts; Solid Deformed Object; Liquid Deformed Object; Object Replacement
14	What is the persuasion of this ad?	Reciprocity; Concreteness : Details about product; Social Impact; Authority/Expert Approval/Third-party approval/Credentials and Awards; Trustworthiness and statistics; Social Identity; Others; Creative; Cheerful; Eager; Active ; Fashionable; Feminine ; Amazed; Emotion; Unclear
15	What is the topic of this ad?	Chocolate; chips; seasoning; Pet food; alcohol; coffee; soda; cars; Electronics; phone, tv, internet,providers; financial; education; security; software; other service; beauty; healthcare; clothing; baby; game; cleaning; home improvement; home appliance; travel; media; sports; shopping; gambling; environment;animal right; human right; safety; smoking,alcohol, abuse; domestic violence; self esteem; political;charities;Unclear;

Overall, We have developed an automated robust pipeline to collect relevant data directly from the web for each ad image both in training and inference. This ensures the availability of rich and diverse contextual data for an ad image without manual intervention. Additionally, for brand-specific information, we tap into publicly available data from company websites and related sources, making it feasible to gather brand-specific annotations effectively.

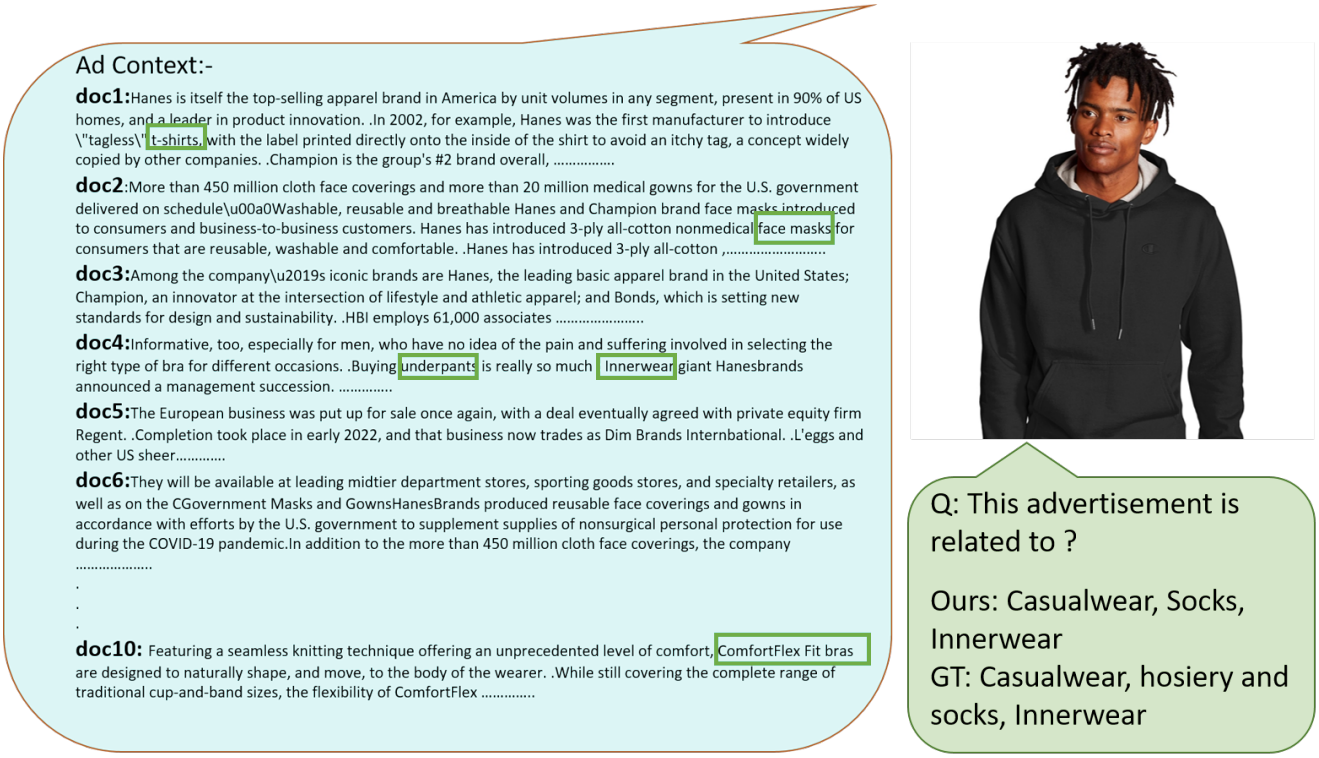

Figure 2. An ad image with ground truth (GT), predicted answer by AdQuestA and question-aware (ranked) ad context

Figure 3. Captions obtained from metadata/BLIP-2 and object labels obtained by Grounding Dino to form a prompt for MIKG module of AdQuestA

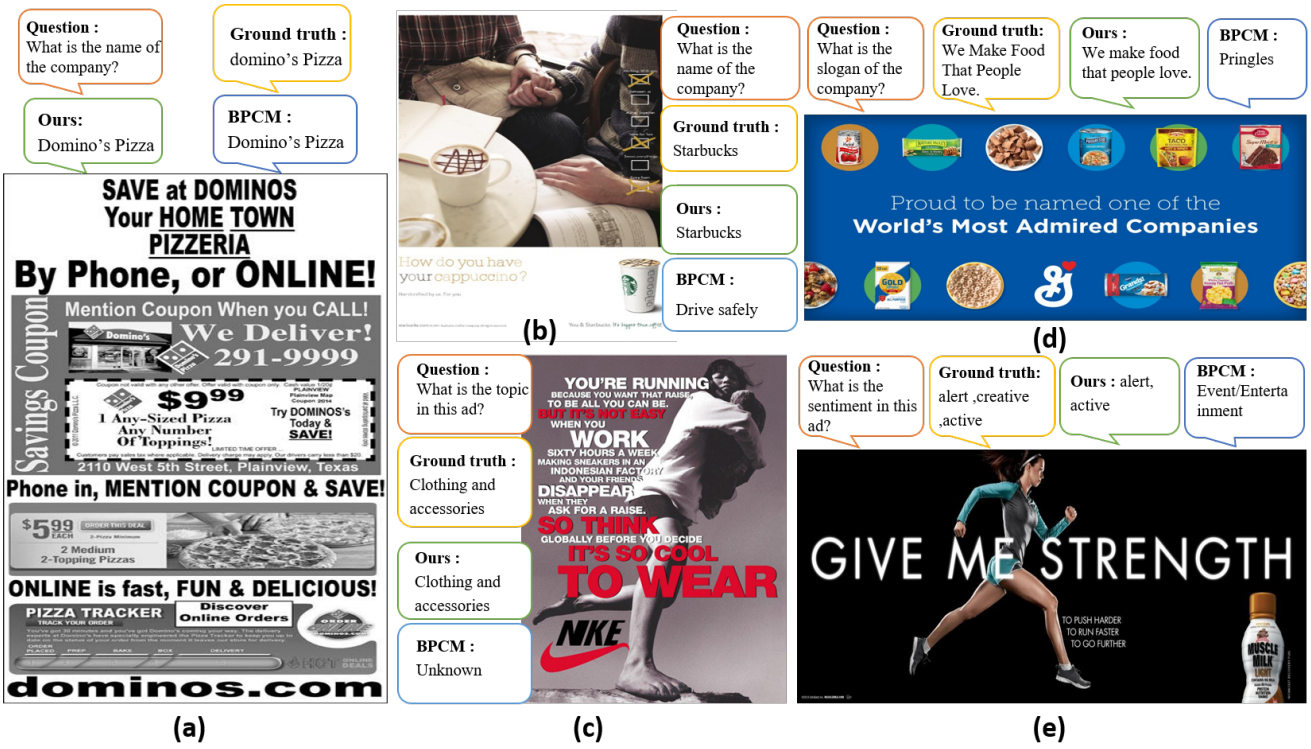

Figure 4. Examples from our ADVQA dataset that AdQuestA generates correct answer compared with the best performing competitive model (BPCM), PICa.

1.2. Appendix B

We employ the pre-trained visual grounding model, Grounding Dino [4], to identify object-centric region proposals. As we need a robust system capable of detecting arbitrary objects specified by human language inputs, thus used a model often referred to as open-set object detection. This model holds significant potential as a generic object detector and finds wide applications across various domains.

Captions for images are extracted from metadata associated with Facebook and Twitter posts and BLIP-2 model if not available in metadata. These regional tags and captions along with visual information are used to form a prompt to feed into LMM to get a implicit knowledge. Example of captions and tags which is use to create a prompt are given in Figure 3.

1.3. Appendix C

We benchmarked our model against various foundational and VQA models. We implemented these models using the same procedures outlined in their respective pipelines, utilizing our ADVQA dataset. All the VQA models with which we compare AdQuestA are as follows:

PICa-BASE [6]: PICa-BASE utilizes VinVL model to generate captions and used them as context with image to fed into GPT-3 to output an answer. It utilises knowledge-based reasoning through a sequence of knowledge retrieval followed by answer prediction.

PICa-Full [6]: PICa-Full uses same pipeline as PICa-Base for context generation but utilises GPT-3 with multi-query ensembling method to predict the improved answer.

KAT [1]: KAT integrates both implicit and explicit knowledge within an encoder-decoder architecture, enabling joint reasoning over these knowledge sources during answer generation. It uses GPT3 (unimodal) to predict implicit knowledge and Wikdata to generate explicit knowledge and then fused in encoder-decoder module. Knowledge Augmented Transformer (KAT) is a different approach that significantly improved the open-domain multimodal VQA task.

These approaches often neglect the relationships within and among object regions, and they sometimes under utilize visual features in the final answering stage. In contrast, REVIVE model emphasizes the importance of leveraging explicit information from object regions throughout both knowledge retrieval and answering stages. It use GLIP model to detect regions. This approach aims to better capture the inherent relationships within object regions, which are crucial for knowledge-based VQA. We did not include the results for this model as its intermediate pre-processing

Figure 5. (A) Success and (B) failure examples obtained by AdQuestA

data files which they fed into encode-decoder module to generate final answer is not publicly available. This preprocessing include a process for getting top-P similar tags and mapping with external knowledge from Wikidata. Therefore, we moved to next best performing model Prophet.

Prophet [5]: This model performs joint reasoning over retrieved knowledge, question and image to predict the answer. First, it generated the candidate keys from the fine-tuned MCAN model which is a vanilla VQA model, and then used these candidate keys with question and caption to

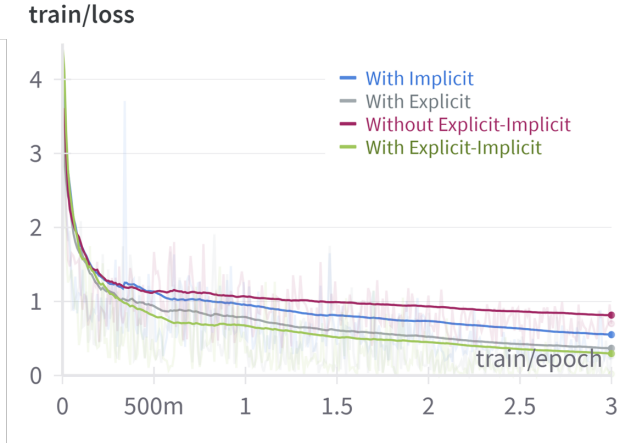

Figure 6. Comparison of training loss per epoch in ablation experiments

help GPT3 to predict improved answer.

We utilise Llama model in place of GPT3 in KAT and Prophet and GPT3 in PICa. PICa is the best performing model among all the models, we compared with, including foundational models GPT-4 etc. BPCM in Figure 4 is representing PICa.

1.4. Appendix D

Here, Figure 4, displays the qualitative results of AdQuestA with that of the best performing competitive model (BPCM), PICa-Full. Details of the figure are given in section 5.6 of the paper.

1.5. Appendix E

We now present success and failure examples obtained by AdQuestA to visualise the performance of the model. Figure 5 (A) has 5 success examples in which AdQuestA achieved accurate answers. In atypicality example, our model predicted one of the ground truth category correctly even though atypicality is present only in 1% images of the dataset. In company name examples, "Burger king" is available as logo in the image whereas "Freddie Mac" might have identified through ad context. Figure 5(B) has a few failure examples where all the predicted answers are close to the ground truth but not exactly matching. In Mcdonald example, the predicted answer is semantically correct and might have been counted in pattern matching. In third example ground truth is "Patients" however answer predicted is "Middle-aged" which is also correct in the sense that the corresponding advertisement is for people more likely to take medicare policies.

References

[1] Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614*, 2021. 4

[2] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1705–1715, 2017. 1

[3] Yaman Kumar, Rajat Jha, Arunim Gupta, Milan Aggarwal, Aditya Garg, Tushar Malyan, Ayush Bhardwaj, Rajiv Ratn Shah, Balaji Krishnamurthy, and Changyou Chen. Persuasion strategies in advertisements. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 57–66, 2023. 1

[4] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 4

[5] Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14974–14983, 2023. 5

[6] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 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, 2022. 4

[7] Eric Zeng, Tadayoshi Kohno, and Franziska Roesner. What makes a "bad" ad? user perceptions of problematic online advertising. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–24, 2021. 1