

# Supplementary

## POP-VQA – Privacy preserving, On-device, Personalized Visual Question Answering

### A. Ablation Study

#### Impact of visual grounding on spatial understanding:

Our initial assumption centered on the potential benefits of visual grounding pre-training for the OFA model in the context of fine-tuning on KVQA data involving face coordinates. During pre-training, OFA uses the bounding box coordinates of objects within the image to align the model to understand spatial relations. We hypothesize that this pre-training is the reason we see marked improvements in POP-VQA performance as compared to earlier works, especially in the questions regarding spatial understanding.

To further bolster our hypothesis, we extended our investigation using a VQA base model that is not pre-trained with bbox coordinates information. Due to limited resources, we could not pre-train OFA architecture from scratch without the bounding box coordinates. Instead, we chose to use a similar architecture based VLM that did not use visual grounding for pre-training. We decided to use the ALBEF [9] architecture for the same. Similar to OFA, ALBEF uses ViT for image encoding, BERT for text encoding and a combination of contrastive and fusion losses to build the multimodal fusion model. ALBEF also has similar number of parameters to OFA, making it a better suited model for comparison. Undoubtedly, it will not be an apple-to-apple comparison. However, we believe the detailed trends to be reflective of the performance variations for PKG-based VQA, on using visual grounding while pre-training.

Employing the same KVQA dataset enriched with meta-

Question Type	LXMERT[K]	ALBEF[K]	POP-VQA[K]
1-hop	48.8	63.4	<b>89.8</b>
Boolean	86.7	88.2	<b>95.7</b>
Comparison	82.5	85.6	<b>89.6</b>
Counting	<b>84.8</b>	82.85	
Intersection	71.5	<b>78.1</b>	73.2
Multi-Entity	73.7	73.2	<b>94.9</b>
Multi-Relation	55.4	50.76	<b>93.27</b>
Spatial	31.1	37.7	<b>83.89</b>
Subtraction	22.2	24.8	<b>37</b>
Overall	52.8	60.1	<b>85.8</b>

Table 6. Performance comparison for visual grounding

data [Section 3.3], we subjected the ALBEF model to same experiments. The results (Table 6) noted under the column **ALBEF(K)** denotes the model performance of this experiment. We also provide a comparison with the results described in the works of Olano et.al [2], who use an LXMERT type VLM architecture to support external knowledge injection. This has been noted in the column marked **LXMERT(K)**. All performance is measured on the KVQA test datasets.

The detailed category-wise results are noted in Table 6. Categories like *Boolean*, *Comparison* and *Intersection* which are more dependent on the core VQA performance itself, see similar performance between ALBEF(K) and POP-VQA(K). The performance seen by LXMERT(K) is lower, as expected given the quality of the VLM used. However, we note a significant decline in performance in *Spatial* category, for both ALBEF(K) and LXMERT(K) - as we had expected given the lack of visual grounding. Both these architectures, fail to learn spatial information as effectively as OFA does. This decrement in spatial accuracy had a cascading effect on the overall accuracy, as evidenced by the findings presented in the accompanying table. These results served to further solidify our conviction that visual grounding pre-training plays a pivotal role in enhancing the outcomes of our experiments.

**Insights to explain Performance Improvement** Our experiments also showcase a significant performance improvement across all sub-categories of data as compared to previous baselines. We account this improvement to two main reasons.

- The chosen baseline VLM is superior in performance to older baselines, having been trained on a larger dataset with more fusion losses to build a deeper aligned space. The OFA pre-training paradigm also uses spatial information, allowing the model to more efficiently generalize to spatial tasks of knowledge based VQA.
- Our chosen training paradigm, allows the model to build a more generalized understanding. Instead of being biased to only one kind of tasks, we teach our

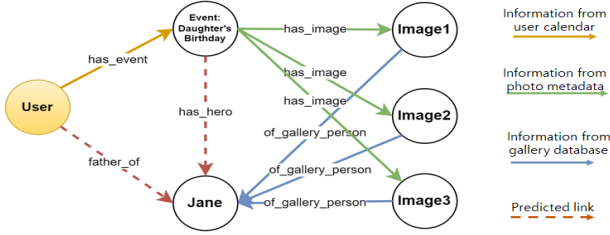


Figure 3. Example Subset from User Knowledge Graph

model to learn when to answer based on its capabilities and when to search and choose for the correct answer from the provided external knowledge.

The results in Table 6 also show that our training paradigm helps extend any generic Vision Language Model to support personalized, KG based answer generation. Our training methodology is able to successfully teach and align the ALBEF model to learn to inject external knowledge effectively. The major loss of accuracy is noted due to the lack of visual grounding. This further validates our methodology.

## B. User Knowledge Graph Creation

In this work, we bridge together the worlds of knowledge graph with visual question answering systems to build a truly personalized user experience. To this regard, we build a completely on-device system that uses information from multiple sources to build a user knowledge graph. We use this area to further describe our system, originally detailed in Section 3.1.

The user knowledge graph is built keeping user as the central node and defines all relevant people/images/events as nodes and the relation as an edge. An example graph subset is shown in Figure 3. Using event information from the calendar (*daughter's birthday*), we build an edge (*Has\_Event*) between user and event node. For this event, through a set of inference rules, we identify that the *hero of event* as *Jane*, who is the *daughter* of the user. Similarly, we identify the related images for this event and add *Has\_Image* relation between the images and the Event node.

Similar to the above example, we build an on-device system that creates a knowledge graph about the user using information from various sources (user profile, calendar, contacts and gallery meta data). Average data distribution statistics, used for PKG, are mentioned in Table 8. These numbers are calculated as an average from all participants in our user study. From the available information, the relevant user/event nodes are created and edges added when possible. Further, using pre-defined rule sets, missing edge information is inferred and updated in the user graph. Table 7 describes the various node definitions, possible edges and inference media.

The underlying on-device storage is built on top of open source graph database called RDF4J. Once hosted on a device, the engine works in the background - collating all the information and building this user personalized KG. When new information is added (new photo clicked, calendar event added, etc.), the KG is updated with relevant nodes and edges, as needed. We then use a SPARQL based query inference system to return relevant information for a selected image. The information is returned in the form of a triple, and used as an input to POP-VQA to generate personalised answers for any user query on a selected image.

## C. User Study

The primary aim of this work is to enhance human interaction with the systems around them. We build an end-to-end pipeline to make the “intelligence” around us, truly personalized and aligned with the user’s life experience. The real test of any such system, however, is with the provided user feedback. The sections below provide in detail the testing environment and experience.

### C.1. Testing Environment

We selected a group of 100 participants using relevant mobile devices (devices that support user KG creation) for the purposes of our study. The test participants were spread equally across both genders, and their age demographics are mentioned in Table 9. All participants were explained the purpose of our solution and application usage before starting. We built an integrated application that allows the user to select an image from their gallery and ask the system

Node	Relation Edge	Inference Source
Person	Person Name	Gallery and Contacts
Person	Relation with user	User Profile, Contacts, Gallery Images, Calendar events
Occasion	Event Information	User Profile, Gallery Images, Calendar events
Hero of Event	Has_Hero (Person)	Calendar Event, Gallery Images, Contacts

Table 7. Examples of Nodes and Inferences for User KG Creation

People (Contacts and Gallery)	Events (Calendar)	Images (Gallery)
1241	3316	30000

Table 8. Average Data Distribution for User Knowledge Graph

Age Group	16-33	34-45	46-70
Percentage of respondents	70%	22%	4%

Table 9. Age Demographics of Test Participants

any question they deemed fit. The application automatically picks up the relevant knowledge graph information for the selected image and uses that for personalized answer generation. The interface showed results from a generic Visual Question Answering system as well as one with personalization layer integrated. Users had the option of clicking on any of the answers to choose which they preferred. They could also mark cases where they felt that neither of the answers were correct or relevant to the asked query.

### C.2. Uniform Testing

Proposed model is built with aim of easing human interactions, and hence a true measure of its efficacy is with how well it translates to real user data. With an aim to measure this metric accurately, we collected real user data from our test participants. During testing process, users could mark which of the two (Generic VQA vs POP-VQA) answers they preferred or mark neither. Users were encouraged to ask 4 – 5 questions per image and ask queries more aligned to how they would interact with smart homes and devices on a daily basis. We collated this information, in the back-end, to attain a total of 5K questions on 1.5K images in total. The performance results on this are noted in Table 4.

### C.3. On-Device Testing

Post system explanation, users were allowed to interact with system freely. As described, they could mark the preferred answer, the details of which we used to understand the difference in user preferences. Users could also click new photos and ask questions about it. This provided an accurate measure of the real-life usage of our solution. At the end, users were asked to fill a survey about their experience. The response to this survey has been noted in Table 5.

## D. Qualitative Analysis

In order to provide a detailed metric of system performance, we showcase various qualitative examples, with user questions and generated answers. All of these examples are taken from user systems with due permissions to circulate their personal data. However, we have blurred out the faces to protect user privacy. We also demonstrate a couple of examples from within our app interface, to further validate our solution.

These examples have been selected to showcase the different media wherein integration of personal knowledge improves the quality and relevance of the answers. As can be seen in R1 and R4, person identification provides more relevant answers than just mentioning generic common nouns. In cases where the user is unable to directly see the image (limited vision or conversing to smart-home systems), such answers provide more useful information to the user. Generic answers of man and woman make absolutely no sense in real-life systems. Similarly, in cases like R6, R7 and R8 - the system capabilities of activity identification becomes more relevant with person identification. R8, especially, looks at a scenario wherein the question itself includes personal information - something that cannot be supported in non-personalized systems. Additionally, location and event identification (R3, R4), inferred from the user knowledge graph, again provides more useful information to the user - helping them remember and reconnect with past events. Knowing the exact event of the image, instead of generic words like "birthday" or "party" provided a more human-like experience to the user - building a system that actually makes human life easier and system interactions seamless.

*Please check the next page for qualitative examples*

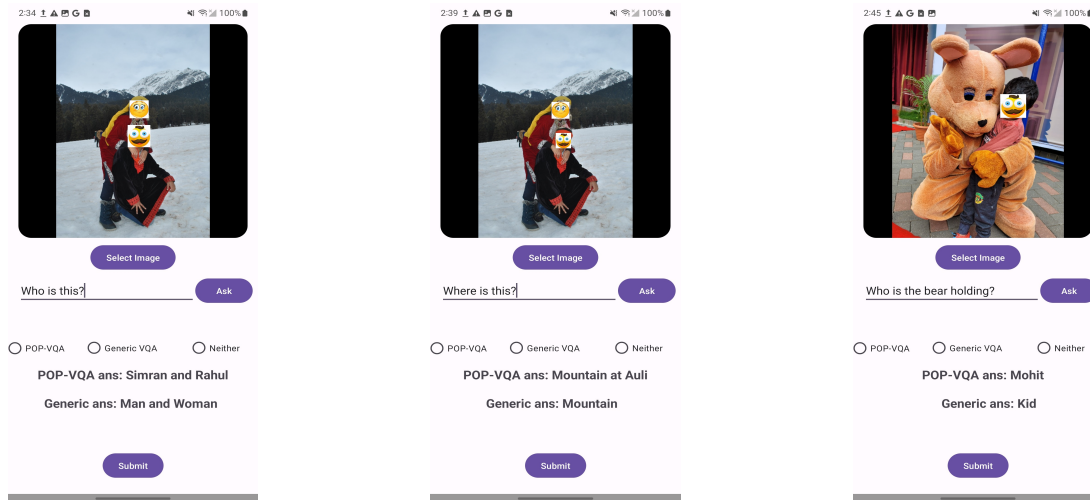










Figure 4. Example scenarios of POP-VQA, on device

1.	Input Image	User Query	Generic VQA answer	POP-VQA answer
2.		Who is in the photo?	Man and woman	Vikram and his wife Sarita
3.		Where was this photo taken?	Beach	Beach at Palolem, Goa
4.		What is happening here?	Birthday	Daughter's birthday party
5.		Whose birthday is it?	Girl	Daughter, Suhali
6.		Who is in the living room?	Boys	Rohan and Mohit
7.		Who is shouting?	Boy	Mohit
8.		Who is making noise?	Girl	Daughter, Rosy
9.		What is Rosy doing?	⟨ Query Not Supported ⟩	Dancing