OVQA: A Dataset for Visual Question Answering and Multimodal Research in Odia Language

Shantipriya Parida[1], **Shashikanta Sahoo**[2], **Sambit Sekhar**[3],**Kalyanamalini Sahoo**[4],
Ketan Kotwal[5], **Sonal Khosla**[3], **Satya Ranjan Dash** [6], **Aneesh Bose**[7],
Guneet Singh Kohli[8], **Smruti Smita Lenka**[3], **Ondřej Bojar**[9]

[1]Silo AI, Finland; [2]Government College of Engineering Kalahandi, India; [3]Odia Generative AI, India;
[4] University of Artois, France; [5]Idiap Research Institute, Switzerland; [6]KIIT University, India;
[7]Microsoft, India; [8]Thapar University, India; [9]Charles University, MFF, ÚFAL, Czech Republic;
correspondence: shantipriya.parida@silo.ai

Abstract

This paper introduces OVQA, the first multimodal dataset designed for visual question-answering (VQA), visual question elicitation (VQE), and multimodal research for the low-resource Odia language. The dataset was created by manually translating 6,149 English question-answer pairs, each associated with 6,149 unique images from the Visual Genome dataset. This effort resulted in 27,809 English-Odia parallel sentences, ensuring a semantic match with the corresponding visual information. Several baseline experiments were conducted on the dataset, including visual question answering and visual question elicitation. The dataset is the first VQA dataset for the low-resource Odia language and will be released for multimodal research purposes and also help researchers extend for other low-resource languages.

1 Introduction

Visual Question Answering (VQA) is a complex task at the intersection of computer vision and natural language processing, requiring models to understand and reason about visual content and formulate accurate responses to textual questions. Despite significant advances in this field, the majority of VQA research has been focused on a handful of widely spoken languages, primarily English. This language bias limits the accessibility and applicability of VQA technologies to non-English speaking populations.

To address this gap, we introduce OVQA, the first multimodal dataset specifically designed for VQA tasks in the Odia language. Odia, an official language of India, is currently spoken by approximately 50 million people.[1] However, it has been largely underrepresented in the realm of natural language processing and VQA research.

By developing a VQA dataset in Odia, we aim to broaden the inclusivity of AI technologies and foster advancements in multilingual and multimodal AI systems.

The OVQA dataset was built by translating 6,149 English question-answer pairs from the widely used Visual Genome dataset into Odia. Each question-answer pair is associated with a unique image, resulting in a robust dataset of 27,809 English-Odia parallel sentences. This ensures a strong semantic alignment between the visual content and the textual data in both languages.

Our contributions are threefold:

- Dataset Creation: We present OVQA, a comprehensive dataset that enriches the multilingual VQA landscape and provides a valuable resource for the Odia language.
- Baseline Experiments: We establish baseline performance metrics through various experiments including visual question answering, and visual question elicitation. These baselines will serve as a reference for future research and development.
- Semantic Alignment: We ensure high-quality translation and semantic consistency between the English and Odia texts, enhancing the dataset's reliability and usability for multimodal learning tasks.

The development of OVQA is a significant step towards bridging the linguistic divide in AI research. By making this dataset publicly available, we hope to inspire further research in multilingual VQA and contribute to the creation of more inclusive AI systems.

For our work, the Visual Genome dataset introduced by Krishna et al. (2016), has been used. It is a large-scale collection of images and associated descriptive data designed to facilitate

---

[1]https://www.britannica.com/topic/Oriya-language

research in computer vision and natural language processing.

We explored the *PaliGemma* (Beyer et al., 2024) model which can be used for various tasks such as VQA, detecting objects on images, or even generating segmentation masks. Here, we have explored the capability of *PaliGemma* for low-resource language on the VQA task. Although *PaliGemma* has zero-shot capabilities – meaning the model can identify objects without fine-tuning, Google strongly recommends fine-tuning the model for optimal performance in specific domains.

2 Related Work

Parida et al. (2023a) created HaVQA, a multimodal dataset for visual question answering for the low-resource Hausa language. The dataset demonstrates several use cases utilizing text and images including multimodal machine translation, visual question answering, and visual question elicitation. Romero et al. (2024) proposed a culturally diverse multilingual Visual Question Answering (CVQA) benchmark which includes culturally driven images and questions from across 28 countries on four continents, covering 26 languages with 11 scripts, providing a total of 9k questions. Gupta et al. (2020) proposed a framework for multilingual and code-mixed VQA for Hindi and English.

3 Focused Language

Odia is an Indo-Aryan language predominantly spoken in Odisha, a state located in eastern India. It is part of the Indo-Aryan language family, which evolved in the Indian subcontinent through three distinct phases: Old Indo-Aryan (1500 BC to 600 BC), Middle Indo-Aryan (600 BC to 1000 AD), and Modern Indo-Aryan (after 1000 AD). Languages that emerged during the Modern Indo-Aryan period include Odia, Bangla, Assamese, Hindi, Urdu, Punjabi, Gujarati, Sindhi, Bhojpuri, Marathi, Sinhali, and Maithili. Odia is thought to have developed around 1000 AD, and it serves as the official language of Odisha, recognized as one of the 22 languages in the Indian constitution. According to the 2011 Census, approximately 42 million people speak Odia. The language features several dialects, with Mughalbandi (Standard Odia) recognized as the standard dialect used in education. The script employed for writing Odia is called the Oriya/Odia script.

3.1 Odia Parts of Speech and Syntax

The primary parts of speech in Odia include nouns, pronouns, verbs, adjectives, and postpositions, along with minor categories such as classifiers, complementizers, and conjunctions (Sahoo, 2001). Odia follows a Subject-Object-Verb (SOV) order, where a simple sentence typically starts with a subject and concludes with a finite verb, placing objects between the subject and the verb, with the indirect object preceding the direct object. Modifiers come before the words they modify: adjectives precede nouns, and adverbs come before verbs. While word scrambling is permitted, the typical structure adheres to V-final patterns, except in poetic contexts.

Example 1:

ମିଲି ମୋତେ ଗୋଟିଏ ବହି ଦେଲ
mili mote goTie bahi delaa
Mili me a book gave
'Mili gave me a book.'

Example 2:

ମିଶିଯାଏ ଯଥା ପ୍ରଭାତୀ ତାରା ରବି କିରଣେ
misijaae jathaa prabhaati taaraa rabi kiraNe
unites as morning star sun ray-PP

ମିଶିଯାଏ ଯଥା ଜୀବାତ୍ମା ପରମାତ୍ମା ଚରଣେ
misijaae jathaa jibaatmaa paramaatmaa charaNe
unites as individual soul great soul of God feet-PP

For instance, Example (1) displays a straightforward sentence, while Example (2) demonstrates poetic inversion, where the verb appears at the beginning of the clause; this inversion is included in our corpus due to the variety of poetic forms.

3.2 Grammatical Features of Odia

Odia features three genders: masculine, feminine, and neuter; two numbers: singular and plural; and eight cases: nominative, vocative, accusative, instrumental, dative, genitive, and locative. There are also three persons: first, second, and third. The subject noun phrase agrees with the verb in terms of person, number, and honorificity. Odia employs a natural gender system, where gender does not influence other grammatical forms like pronouns or verbs. Although gender is explicitly marked in nouns and adjectives, pronouns do not show overt gender distinctions; they are generally neutral. The gender of a pronoun is determined

by the noun or adjective it associates with (Parida et al., 2023b). In Odia, there is a four-fold tense distinction: past, present, future, and hypothetical, based on whether an event occurs before, during, after, or in a hypothetical context. The present tense marker is not morphologically expressed, while the other three are indicated by -il (past), -ib (future), and -ant (hypothetical) (Sahu et al., 2022; Parida et al., 2020; Nayak, 1987).

4 Odia VQA Dataset

In this section, we delve into the various stages involved in the OdiaVQA dataset creation process, including collection, annotation, validation, and data analysis.

4.1 Data Collection and Annotation

For the creation of new dataset, we utilized the Visual Genome Dataset[2] as our primary source of images, supplemented with question-answer pairs. This dataset offers a rich multimodal context comprising images and relevant captions. To gather data for the VQA task, we developed a specific web interface. With the assistance of seven native Odia speakers to manually translate the QA pairs, we annotated the dataset via this web interface.

The interface was thoughtfully designed to integrate an Odia keyboard as shown in Fig. 1, facilitating easy access to special characters in Odia. A detailed guideline was provided to the annotators to minimize errors during the annotation process. Notably, annotations were not supposed to be generated using translation tools, and annotators were required to view the images before annotating the QA pairs. These measures were implemented to reduce errors in annotations, ensuring the authenticity and overall quality of the dataset.

4.2 Data Validation

Concurrently with the annotation process, each question-answer pair underwent validation to ensure translation consistency and quality. The validation process included basic spelling and grammar checks using the interface. We engaged seven native Odia speakers to validate the entire dataset simultaneously with the annotation process. A separate interface was employed for the

validation process that simultaneously displayed images and translated question-answer pairs to the validators. Validators could update question-answer pairs in case of errors, and any changes made were directly reflected in the back-end as well.

Item	Count
Number of Images	6,149
Number of Questions	27,809
Number of Answers	27,809
Number of Wh-Questions	26,939
Number of Counting Questions	70
Others	800

Table 1: Statistics of the OVQA Dataset.

4.3 Data Analysis

Within the OVQA dataset, the Odia Natural Language Processing (ONLP)[3] toolkit alongside a Basic Tokenizer has been employed for Odia text tokenization. Table 1 presents pertinent statistics, Question and answer length in OVQA dataset is shown in Figure 2.

Odia	Gloss	Percentage (%)
କଣ	What	56.67
କେଉଁଠାରେ	Where	16.30
କିପରି	How	12.83
କିଏ (କାହାର)	Who (whose)	6.04
କେବେ	When	5.24
କାହିଁକି	Why	3.15

Table 2: Statistics of the OVQA Dataset based on the Question Types.

4.4 Question

In the original English dataset, there are various question types, which can be classified into two main categories: wh-questions and counting questions. Wh-questions typically begin with words such as 'What,' 'Where,' 'When,' 'Whens,' 'Who,' 'How,' and 'Whose.' The statistics for different types of *wh*-questions are presented in Table 2. Odia questions range from as short as two words to as long as eleven words. The distribution of question lengths is illustrated in Fig. 3.

---

[2]https://homes.cs.washington.edu/~ranjay/visualgenome/index.html

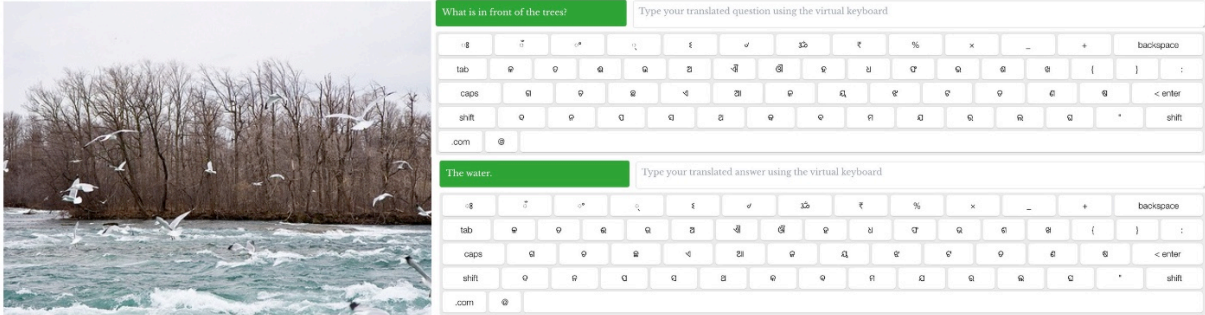[3]https://github.com/nlpodisha/oriya-nlp

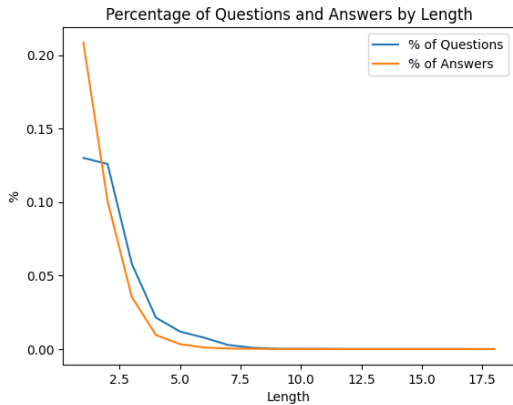Figure 1: Odia Visual Question Answer (OVQA) Annotation Interface

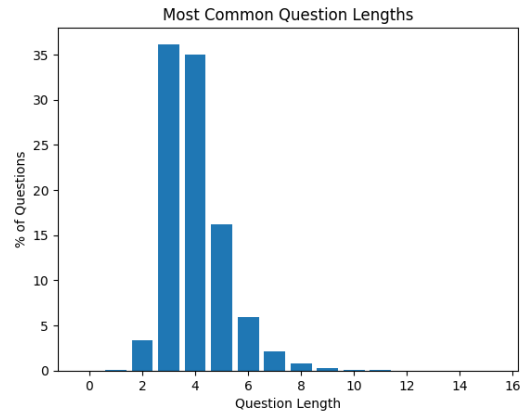Figure 2: Percentage of Questions and Answers by length.

Figure 3: Distribution of Question Length.

4.5 Answers

Depending on the questions in OVQA dataset, different lengths of answers are included. In the majority of the cases (60% cases out of more than 20k QA pairs), the shortest answer is just one word or just a number; however, the longest answer is eight words. The distribution of the length of the answers is shown in Fig. 4 for different types of questions.

5 Baselines for Use Cases

5.1 Visual Question Answering

We used *PaliGemma-3b-448mix* [4] from Google for model fine-tuning on the VQA task. *PaliGemma* is a 3B vision-language model composed of a SigLIP vision encoder and a Gemma language decoder linked by a multimodal linear projection (Beyer et al., 2024; Fedorov et al., 2022).

We prepared the dataset into an instruction set format for fine-tuning.

We used Deepspeed[5] for training on GPU. For GPU, we used AMD Instinct MI250X Accelerator where each node has 60GB GPU memory and we have 1*8 nodes.

We used supervised fine-tuning (SFT) for the full fine-tuning. The hyperparameters are shown in Table 3 and learning curve in Fig. 6.

Hyper Parameter	Value
Train Batch Size (per device)	2
Gradient Accumulation Steps	4
Warm-sup step	50
Learning Rate	$3e^{-4}$
LR_Scheduler	Cosine
Epochs	10
Cutoff Length	1536
bf16	True

Table 3: Training Hyperparameters for VQA

5.2 Visual Question Elicitation

We used the images and associated questions to train an automatic VQE model (Fedorov et al., 2022). We extracted visual features using the images and fed them to an LSTM decoder.

---

[4]https://huggingface.co/google/paligemma-3b-mix-448

[5]https://github.com/microsoft/DeepSpeed

The decoder generates the tokens of the caption autoregressively using a greedy search approach (Soh, 2016). Trained to minimize the cross-entropy loss on the questions from the training data (Yu et al., 2019a) was minimized.

Image encoder All the images were resized to 224×224 pixels, and features from the whole image were extracted to train the model. The feature vector is the output of the final convolutional layer of ResNet-50. It is a 2048-dimensional feature representation of the image. The encoder module is a fixed feature extractor and, thus, non-trainable.

LSTM decoder A single-layer LSTM, with a hidden size of 256, was used as a decoder. The dropout is set to 0.3. During training, for the LSTM decoder, the cross-entropy loss is minimized and computed using the output logits and the tokens in the gold caption. Weights are optimized using the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.001. Training is halted when the validation loss does not improve for ten epochs. We trained the model for 100 epochs.

6 Discussion and Analysis

6.1 Visual Question Elicitation

Since it is challenging to assess the quality of the generated questions using automatic evaluation metrics, we conducted a manual evaluation with the assistance of a native Odia speaker. Around 10% of the generated questions were sampled and manually reviewed. Each question was categorized as 'Exact,' 'Correct,' 'Nearly Correct,' or 'Wrong.' The distribution of these categories
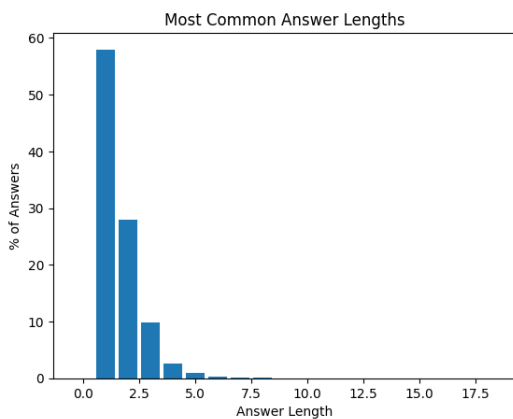

Figure 4: Distribution of Answer Length.

[{ "content": "ଫଟୋରେ କେତୋଟି ଜେବ୍ରା ଅଛି?", "role": "user" }, { "content": "୨ |", "role": "assistant" }]
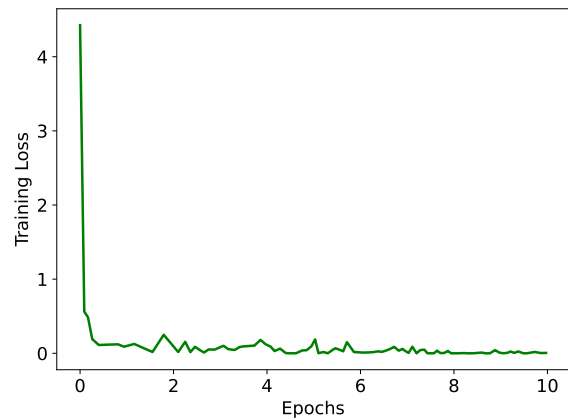
Figure 5: Dataset Sample

Figure 6: Learning Curve for VQA Training

is shown in Figure 8, with additional sample questions provided in Part A in the Appendix.

All the generated predictions were valid and reasonable questions, with 99.5% of them (all but 3) correctly ending with a question mark ('?'). The distribution of question types is as follows: "କଣ" (what)–60.1%, "କେଉଁଠାରେ" (where)–26.4%, "କେବେ" (when)–4.8%, "କିଏ" (who)–3.7%, "କାହିଁକି" (why)–1.5%, "କେତେ" (how much)–2.32%, and "କିପରି" (how)–1.2%.

7 Availability

The OVQA dataset can be accessed via LINDAT at: http://hdl.handle.net/11234/1-5820.

Additionally, the OVQA dataset, designed for multimodal LLM training in an instruction set format, is available on Hugging Face:

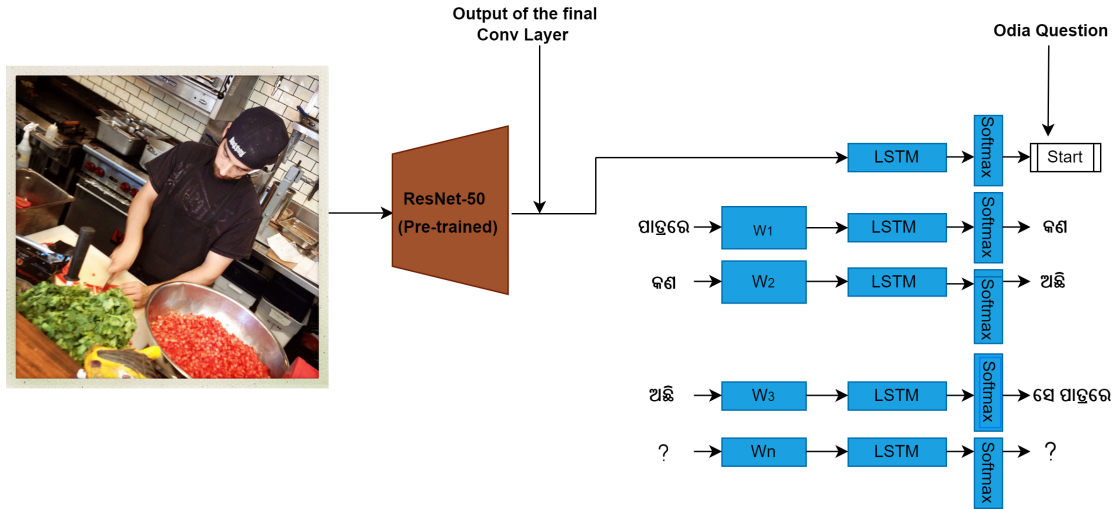Dataset: https://huggingface.co/datasets/odiagenmllm/odia_vqa_en_odi_set

Figure 7: Architecture of Visual Question Elicitation using ResNet-50 (Koonce and Koonce, 2021) and LSTM (Yu et al., 2019b). The training question was "ପାତ୍ରରେ କଣ ଅଛି ?" (gloss: What is in the bowl?). During inference, when the image was passed through the system, the generated question was "କଣ ଅଛି ସେପାତ୍ରରେ?" (gloss: What is in the container?).
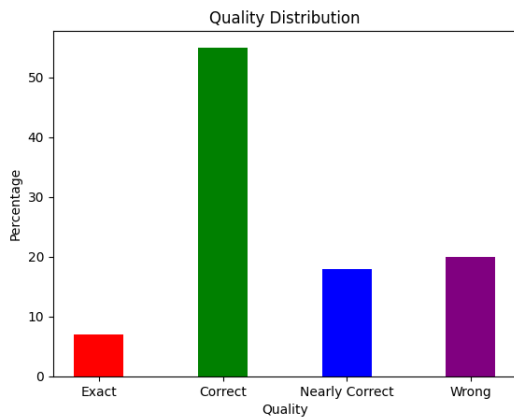

Figure 8: Quality Distribution of Automatically Generated Questions.

8 Conclusion

In this work, we presented **OVQA**: a multimodal dataset suitable for various NLP tasks for the Odia language. Some examples of these tasks include VQA, VQE, and other research tasks based on multimodal analysis.

The OVQA dataset is available for research and non-commercial use under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License.[6]

Our planned future work includes: *i)* extending the dataset with more images depicting regional, and cultural aspects and QA pairs *ii)* providing ground truth for all images for image captioning experiments, and *iii)* organizing a shared task using the OVQA dataset.

Ethics Statement

We do not envisage any ethical concerns. The dataset does not contain any personal, or personally identifiable, information, the source data is already open source, and there are no risks or harm associated with its usage.

Limitations

The most important limitation of our work lies in the size of the OVQA dataset. However, substantial further funding would be needed to resolve this.

Acknowledgements

References

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

[6]https://creativecommons.org/licenses/by-nc-sa/4.0/

Dmitry A Fedorov, Bo Peng, Niranjan Govind, and Yuri Alexeev. 2022. Vqe method: a short survey and recent developments. *Materials Theory*, 6(1):2.

Deepak Gupta, Pabitra Lenka, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A unified framework for multilingual and code-mixed visual question answering. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 900–913, Suzhou, China. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Brett Koonce and Brett Koonce. 2021. Resnet 50. *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*, pages 63–72.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Preprint*, arXiv:1602.07332.

Rath Nayak. 1987. *Non-finite clauses in Oriya*. Doctoral dissertation, Central Institute of English and Foreign Languages (CIEFL), Hyderabad, India.

Shantipriya Parida, Idris Abdulmumin, Shamsuddeen Hassan Muhammad, Aneesh Bose, Guneet Singh Kohli, Ibrahim Said Ahmad, Ketan Kotwal, Sayan Deb Sarkar, Ondřej Bojar, and Habeebah Kakudi. 2023a. HaVQA: A dataset for visual question answering and multimodal research in Hausa language. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10162–10183, Toronto, Canada. Association for Computational Linguistics.

Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2020. Odiencorp: Odia–english and odia-only corpus for machine translation. In *Smart Intelligent Computing and Applications: Proceedings of the Third International Conference on Smart Computing and Informatics, Volume 1*, pages 495–504. Springer.

Shantipriya Parida, Alakananda Tripathy, Satya Ranjan Dash, and Shashikanta Sahoo. 2023b. Mdolc: Multi dialect odia song lyric corpus.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*.

Kalyanamalini Sahoo. 2001. *Oriya Verb Morphology and Complex Verb Constructions*. Ph.d dissertation, Norwegian University of Science and Technology, Trondheim, Norway.

Anupama Sahu, Sarojananda Mishra, and Kalyan Kumar Jena. 2022. Classification of odia and other text printed images usingmachine intelligence based approach. *NeuroQuantology*, 20(9):764.

Moses Soh. 2016. Learning cnn-lstm architectures for image caption generation.

Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. 2019a. Multimodal transformer with multi-view visual representation for image captioning. *IEEE transactions on circuits and systems for video technology*, 30(12):4467–4480.

Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. 2019b. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270.

A Visual Question Elicitation Sample Predictions

Example 1: Exact	Example 2: Correct
Ref Que.: ସେଠାରେ କେତେଗୁଡ଼ିଏ ଫ୍ୟାକେଟ୍ ଅଛି? **Gloss:** How many faucets are there? **Pred Que.:** ସେଠାରେ କେତେ ଫ୍ୟାକେଟ୍ ଅଛି? **Gloss:** How many faucets are there?	**Ref Que.:** ଆକାଶରେ ମେଘର କେଉଁ ରଙ୍ଗ? **Gloss:** What color are the clouds in the sky? **Pred Que.:** ମେଘର ରଙ୍ଗ କ'ଣ? **Gloss:** What is the color of the clouds?
Example 3: Nearly Correct	**Example 4: Wrong**
Ref Que.: ବିମାନରେ ଥିବା ଫିନ୍ ଉପରେ ଅକ୍ଷରଗୁଡ଼ିକ କ'ଣ? **Gloss:** What are the letters on the fin on the airplane? **Pred Que.:** ବିମାନରେ ଅକ୍ଷରଗୁଡ଼ିକ କ'ଣ? **Gloss:** What are the letters on the airplane?	**Ref Que.:** ସବୁଜ ପରିବାଟି କଣ? **Gloss:** What is the green vegetable? **Pred Que.:** ବ୍ରୋକଲି ଟିକେଉଁଠାରେ ଅଛି? **Gloss:** Where is the broccoli?

Table 4: Visual Question Elicitation Sample Predictions

B Recruitment of Annotators and Validators

We selected native Odia speakers from the Odia Generative AI (OdiaGenAI) research group, which consists of experienced translators, to serve as annotators and validators. The annotation team comprised 4 women and 3 men, while the validation team included 3 women and 4 men. Each team member holds at least an undergraduate degree and resides in various regions across Odisha state of India.

C Annotation Guidelines

The following instructions were provided to the Odia annotators and validators:

1. Review the Odia typing guidelines carefully. Before beginning the annotation, perform a quick test and report any issues encountered.
2. Ensure that the annotator is a native speaker of the Odia language.
3. View the image before proceeding with annotation.
4. Aim to understand the task fully—translate both questions and answers into Odia.
5. Refrain from using any machine translation tools for annotation.
6. Do not enter dummy entries for testing the interface.

7. Data will be saved at the backend.
8. Press the Shift Key on the virtual keyboard for complex consonants.
9. Contact the coordinator for any clarification/support.