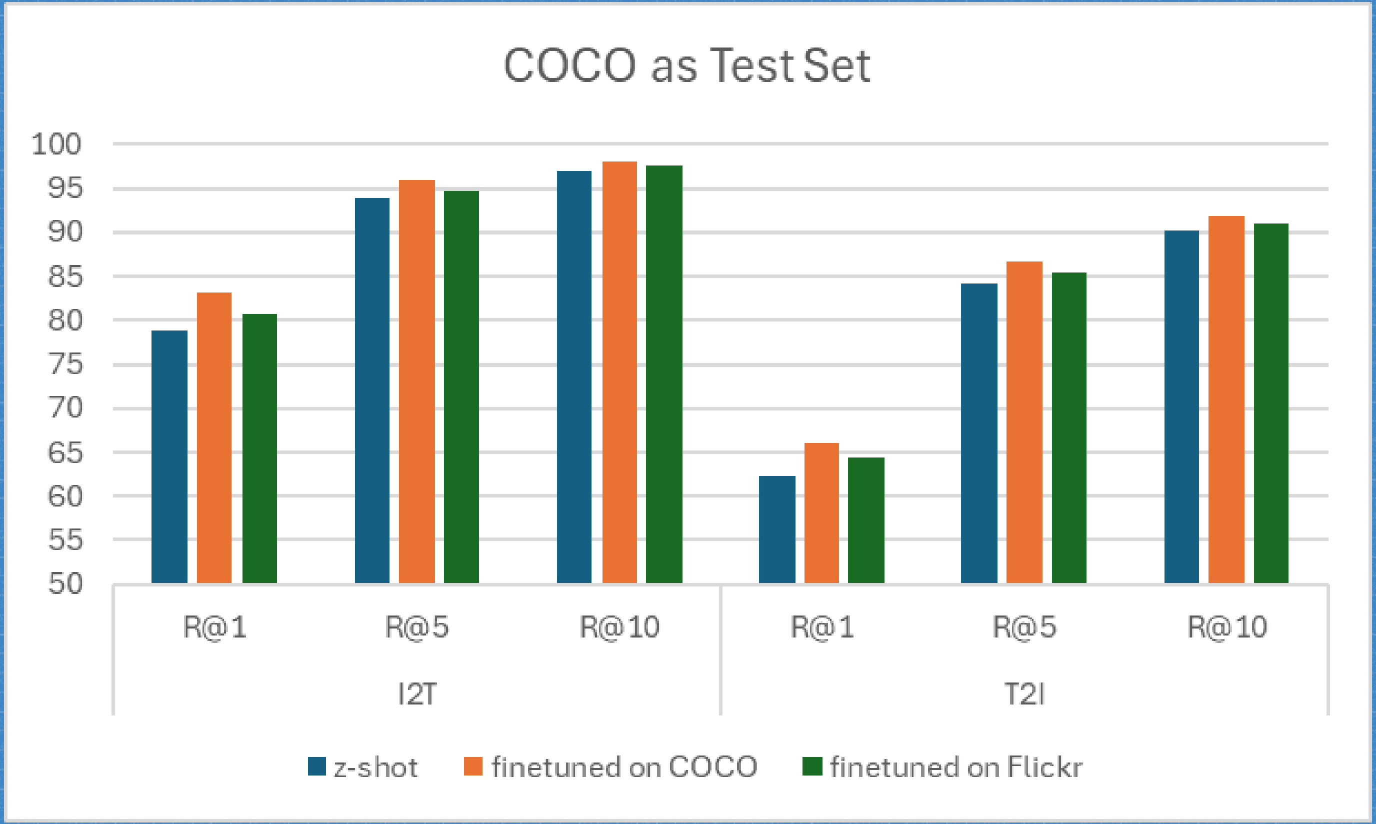
WikiDO

A New Benchmark Evaluating Cross-Modal Retrieval for Vision-Language Models

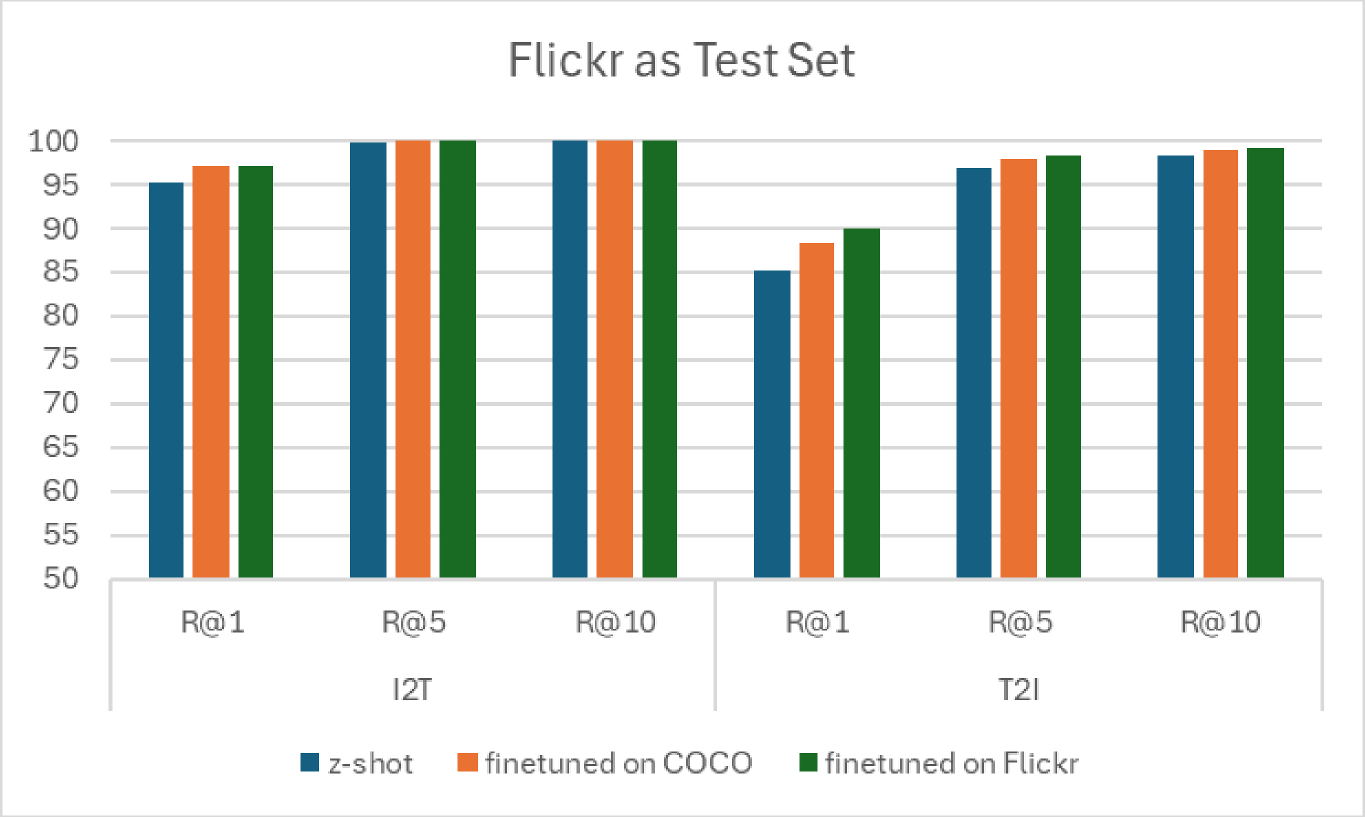
Pavan Kalyan Tankala*, Piyush Singh Pasi*,
Sahil Dharod, Azeem Azaz Motiwala,
Preethi Jyothi, Aditi Chaudhary, Krishna Srinivasan

IIT Bombay

DeepMind

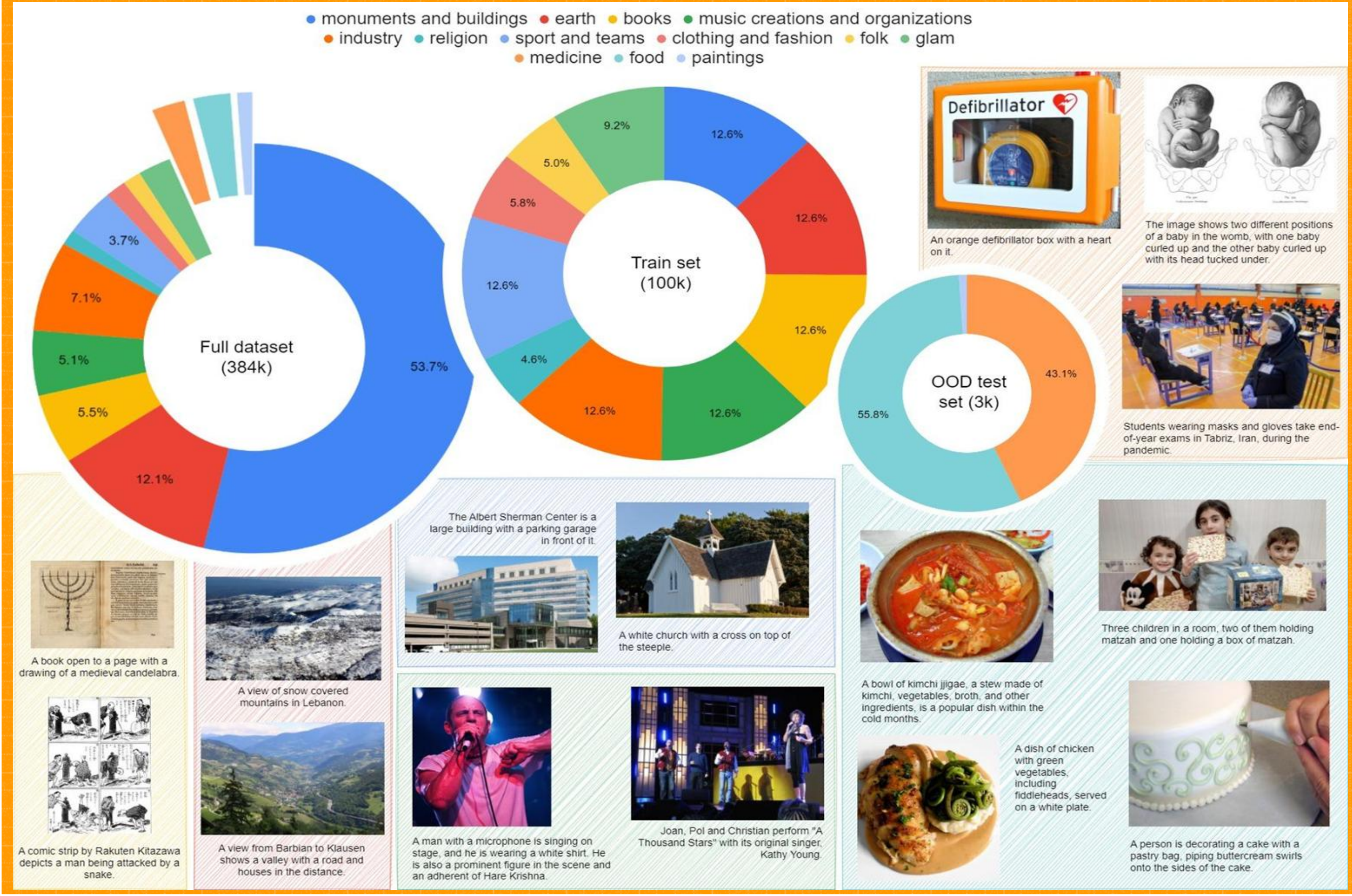


Current Method of testing VLMs (BLIP) for OOD generalization



Current Method of testing VLMs (BLIP) for OOD generalization

WikiDO



An orange defibrillator box with a heart on it.

The image shows two different positions of a baby in the womb, with one baby curled up and the other baby curled up with its head tucked under.

Students wearing masks and gloves take end-of-year exams in Tabriz, Iran, during the pandemic.

A book open to a page with a drawing of a medieval candelabra.

A comic strip by Rakuten Kitazawa depicts a man being attacked by a snake.

A view of snow covered mountains in Lebanon.

A view from Barbian to Klausen shows a valley with a road and houses in the distance.

The Albert Sherman Center is a large building with a parking garage in front of it.

A white church with a cross on top of the steeple.

A man with a microphone is singing on stage, and he is wearing a white shirt. He is also a prominent figure in the scene and an adherent of Hare Krishna.

Joan, Pol and Christian perform "A Thousand Stars" with its original singer, Kathy Young.

A bowl of kimchi jjigae, a stew made of kimchi, vegetables, broth, and other ingredients, is a popular dish within the cold months.

A dish of chicken with green vegetables, including fiddleheads, served on a white plate.

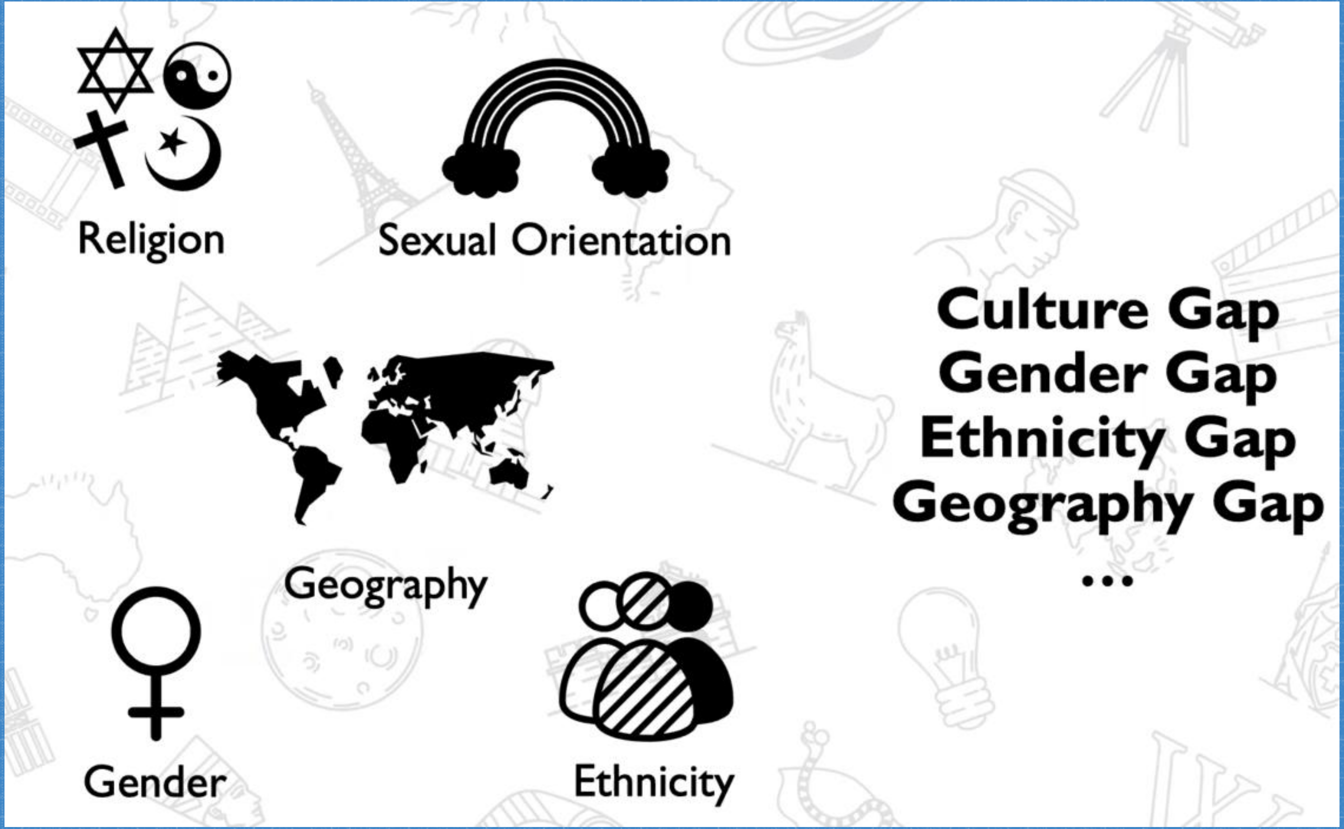
Three children in a room, two of them holding matzah and one holding a box of matzah.

A person is decorating a cake with a pastry bag, piping buttercream swirls onto the sides of the cake.

- We curate a new dataset WikiDO consisting of 1) 354K training images with corresponding text and 2) two evaluation sets -- an in-domain (ID) set and an out-of-domain (OOD) set

- WikiDO spans different topics such as food, books, fashion and sports.

WikiDO is derived from the Wikipedia Diversity Observatory



Source: Wikipedia Diversity Observatory - Meta

Data Curation

- We extract URLs of images in Wikipedia pages from Wikipedia dumps

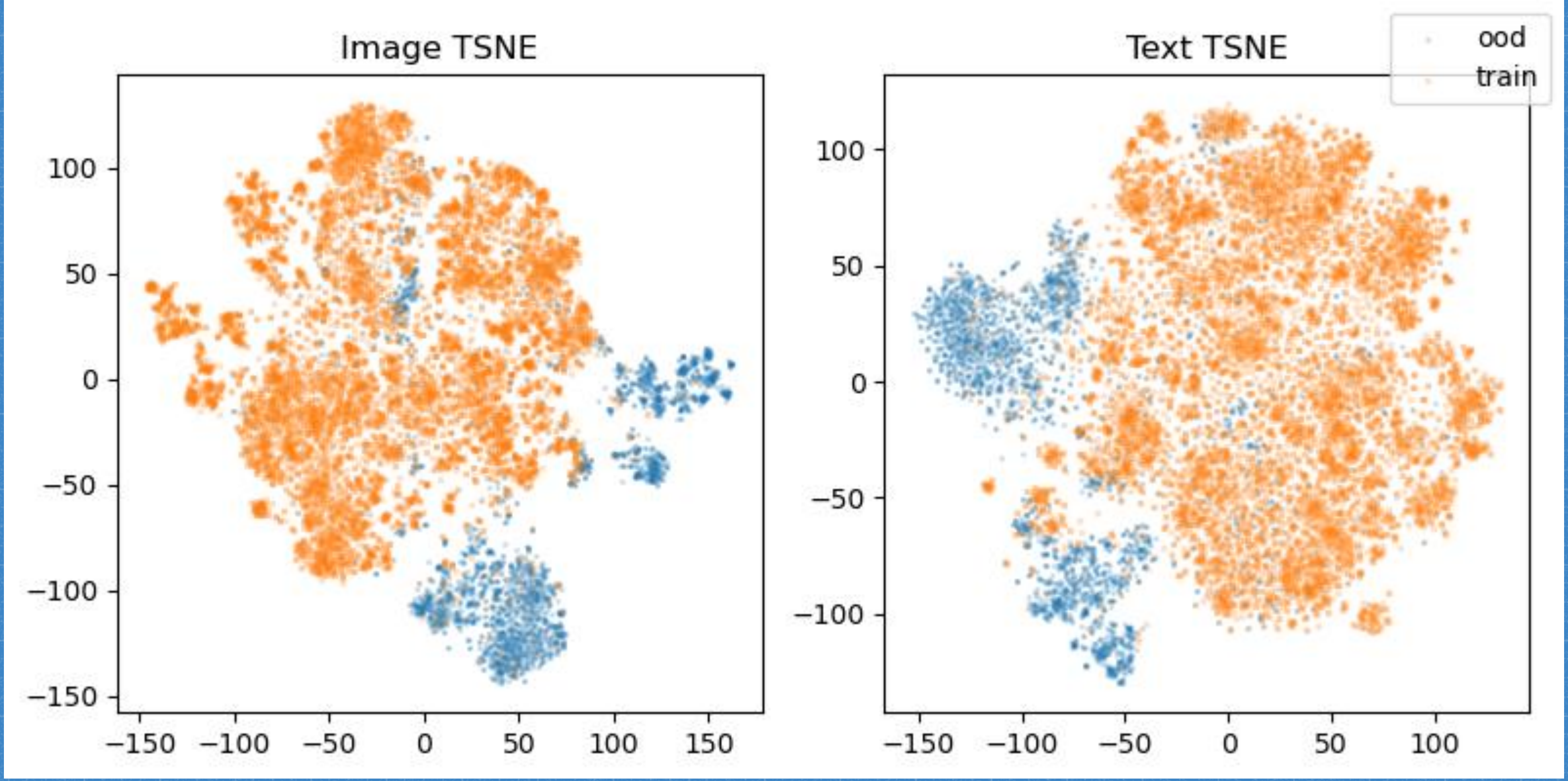
- All images of the page are assigned topic of that page

- We extract 1.2M unique images and its metadata, filter it using WIT method which results in 384K image-text pairs with topic labels.

- Image Captions were enhanced using LLAVA via prompting with both image and current caption.

Field	Description
image_path	Path of the image
image_id	Wiki ID of the image
orig_cap	Reference text from Wikipedia
image	Unique image ID given in the dataset
page_id	Wiki ID of the page from which the image was extracted
page_title	Title of the Wikipedia article from which the image was extracted
topic	Topic label from Wikipedia Diversity Observatory
caption	Caption obtained by passing orig_cap through LLava (human verified for test, val sets)

Domain Gap and Test split



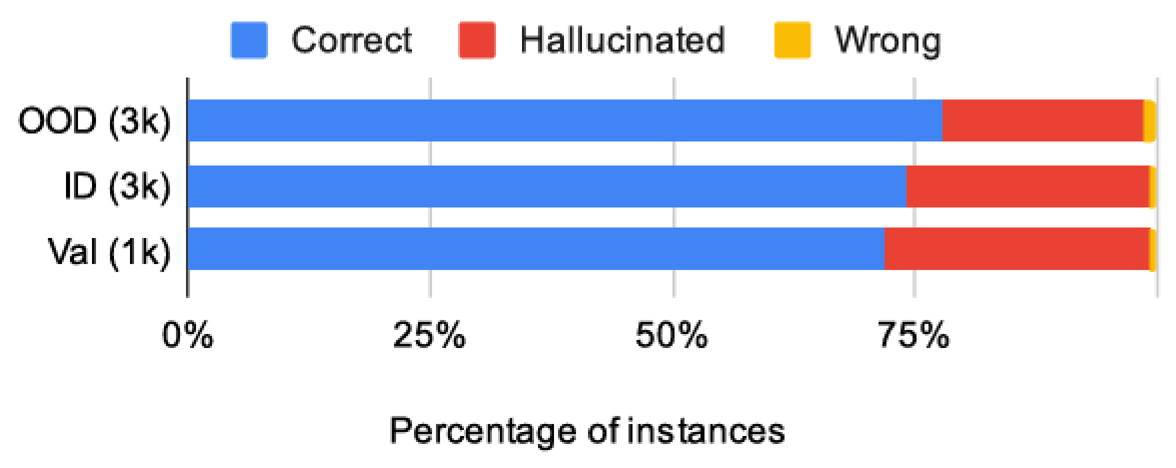
Human Verification

As captions are enhanced by LLAVA it can hallucinate
Hence, human verification was done as follows:

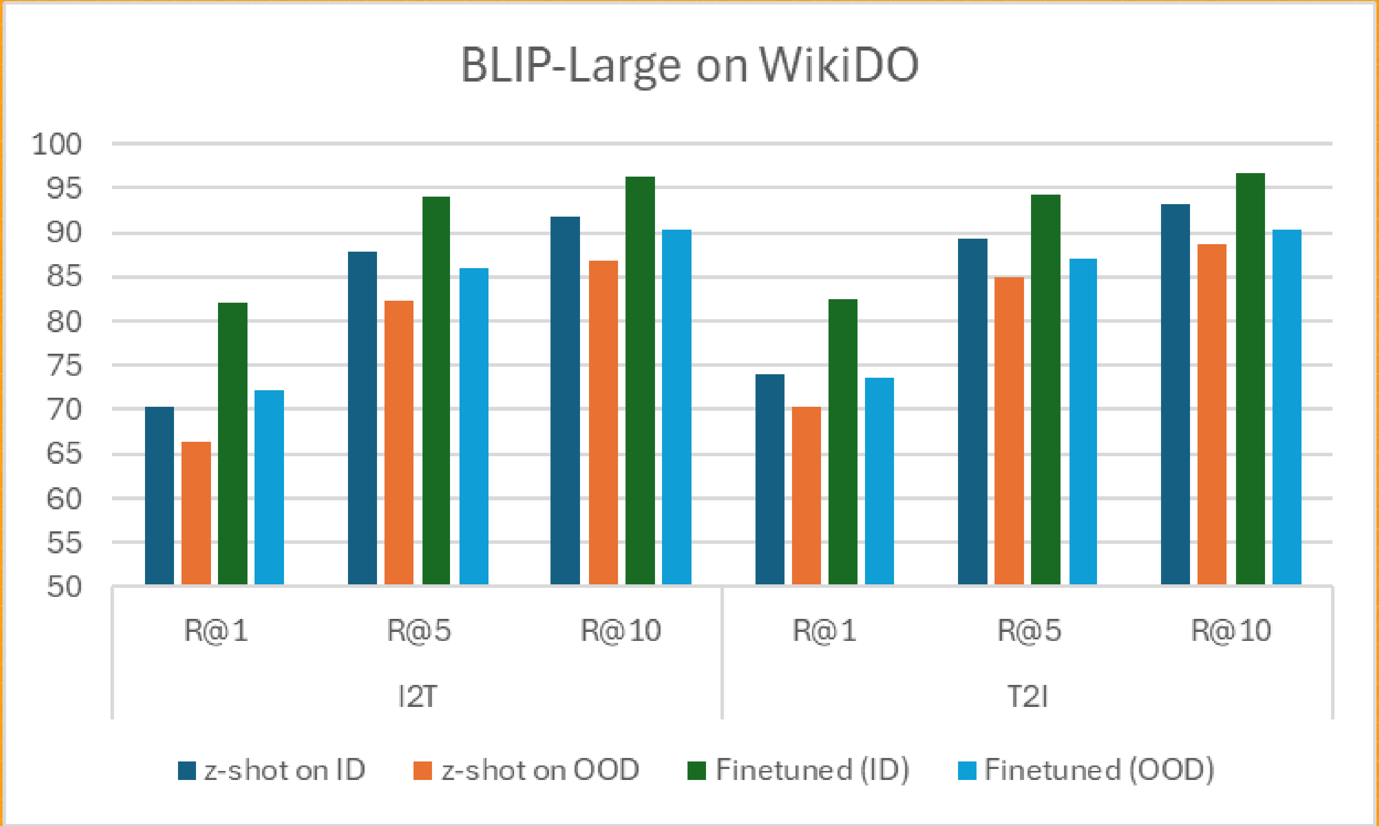
Is there any made-up/ hallucinated content in the caption that is not supported by the image/reference text?

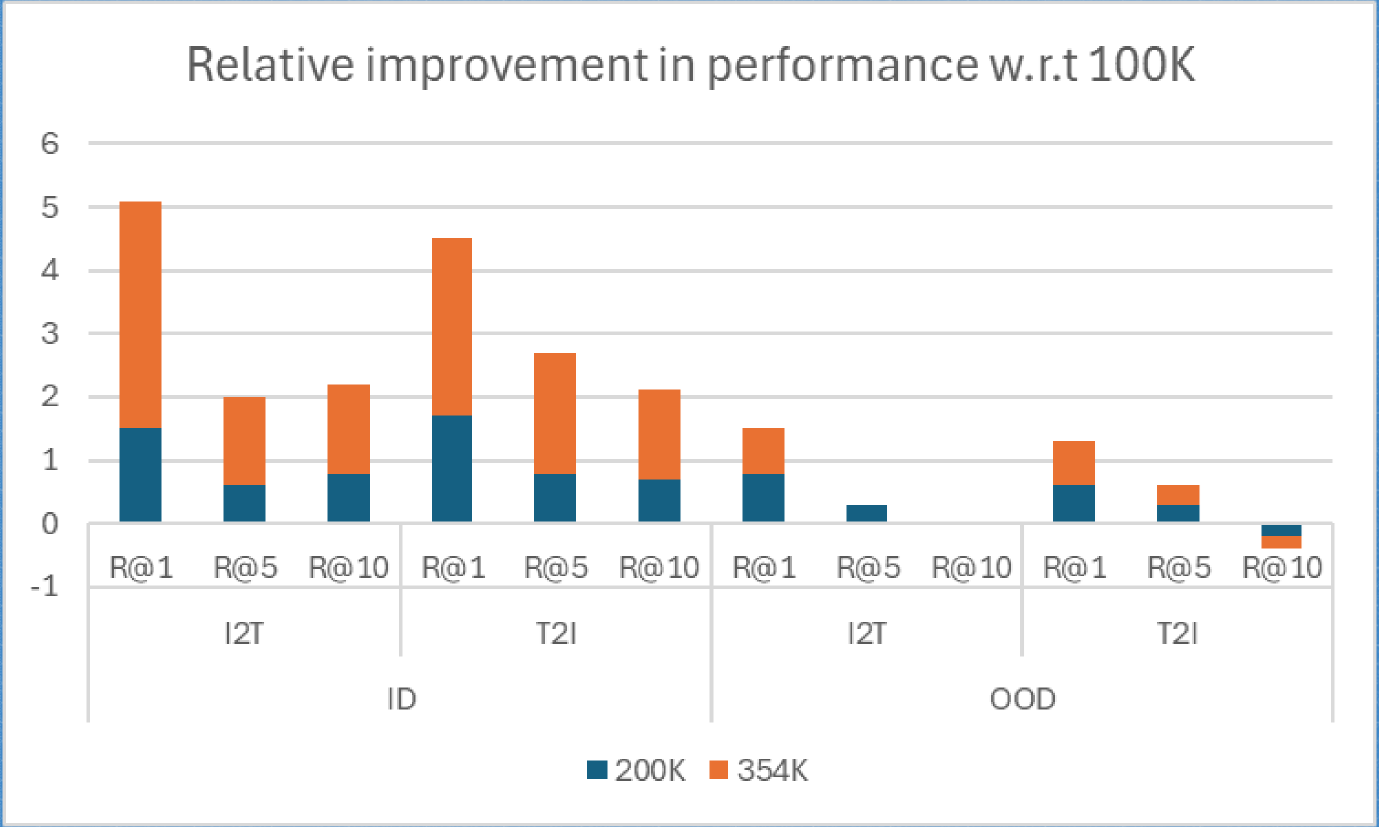
1. Yes
2. No

If Yes, correct the reference text by mainly removing the hallucinations.

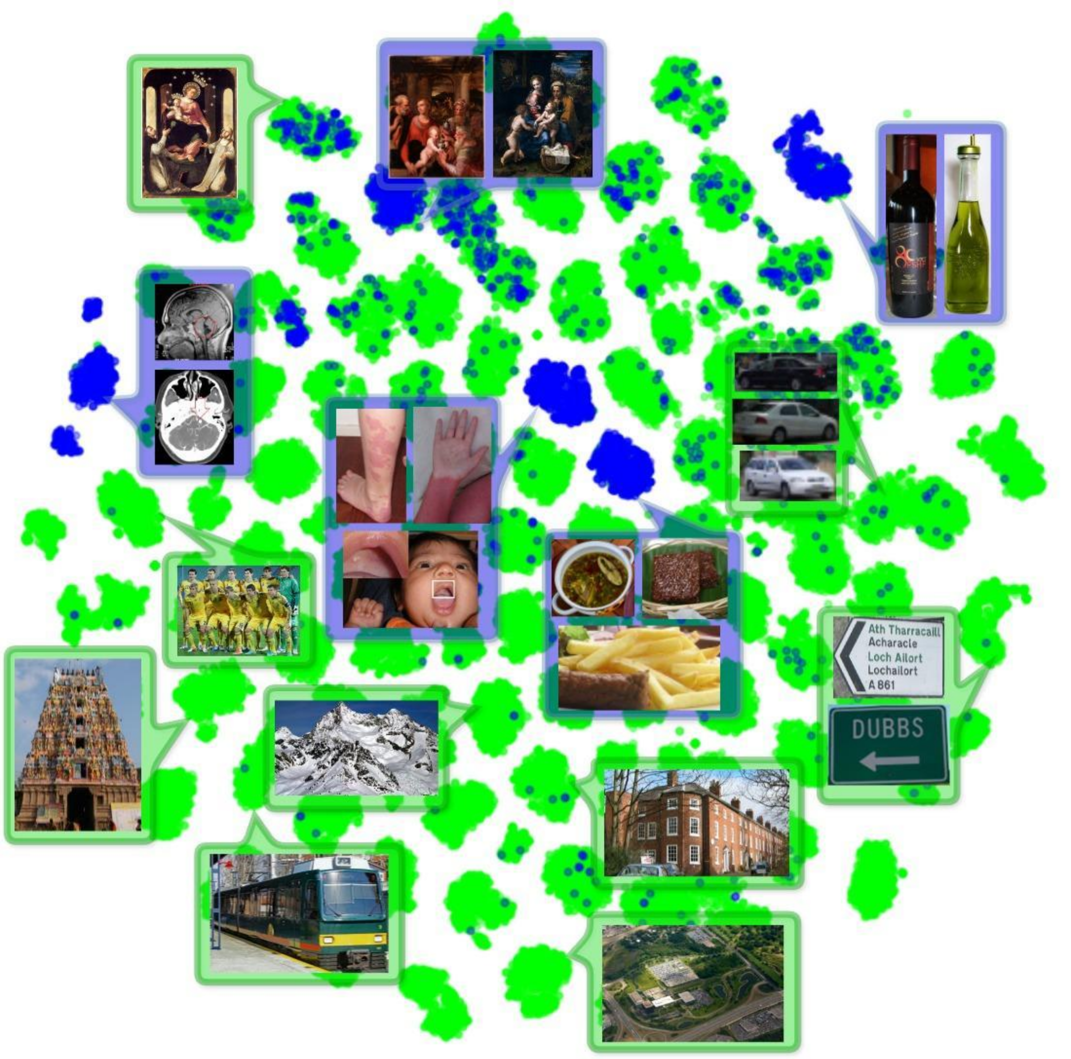


Results on WikiDO Dataset





Minimal improvements in OOD, suggesting that scaling ID data is insufficient to close the performance gap



TSNE of 100 object clusters- OOD objects and ID objects

While there are a few clearly separated clusters for OOD objects, there are clusters that contain both objects in OOD and ID instances. This object overlap explains the gains in R@K for OOD after fine-tuning.

Thank you

Link to benchmark: https://www.kaggle.com/competitions/wikido24/