

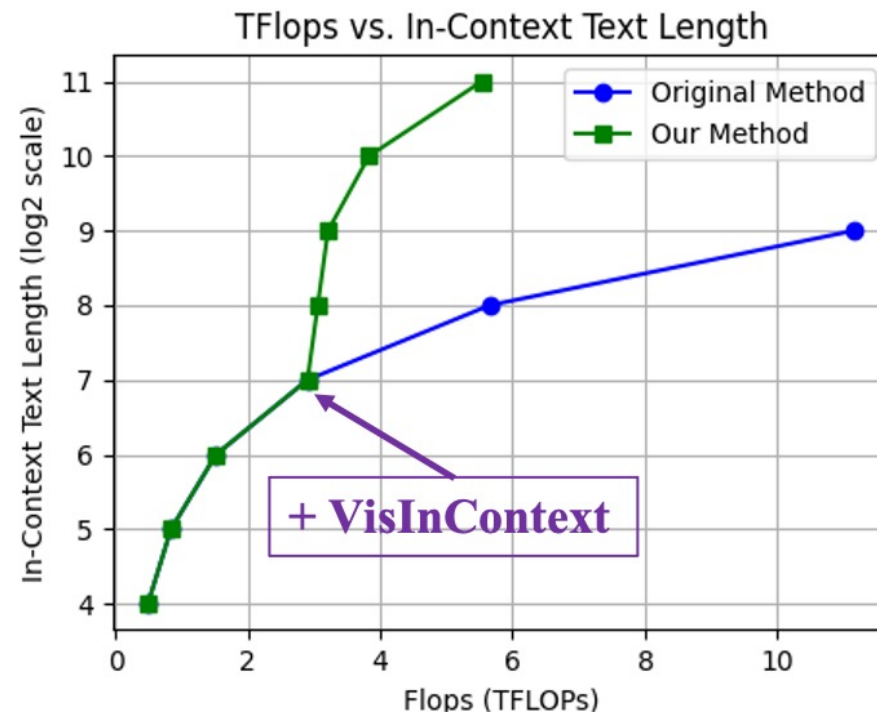
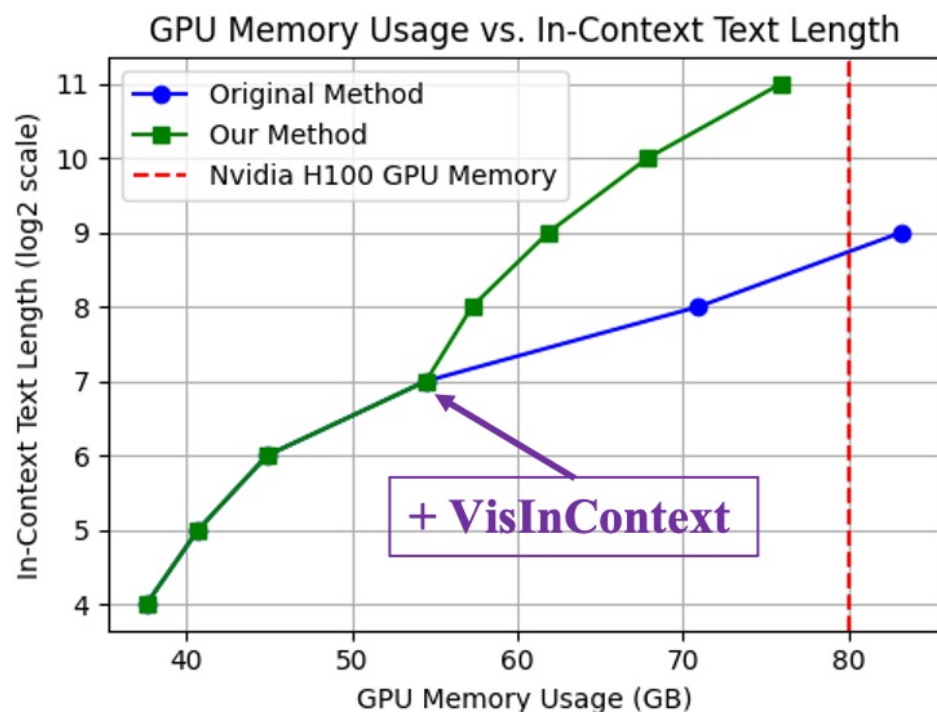


Leveraging Visual Tokens for Extended Text Contexts in Multi – Modal Learning

Presented by [Alex] Jinpeng Wang

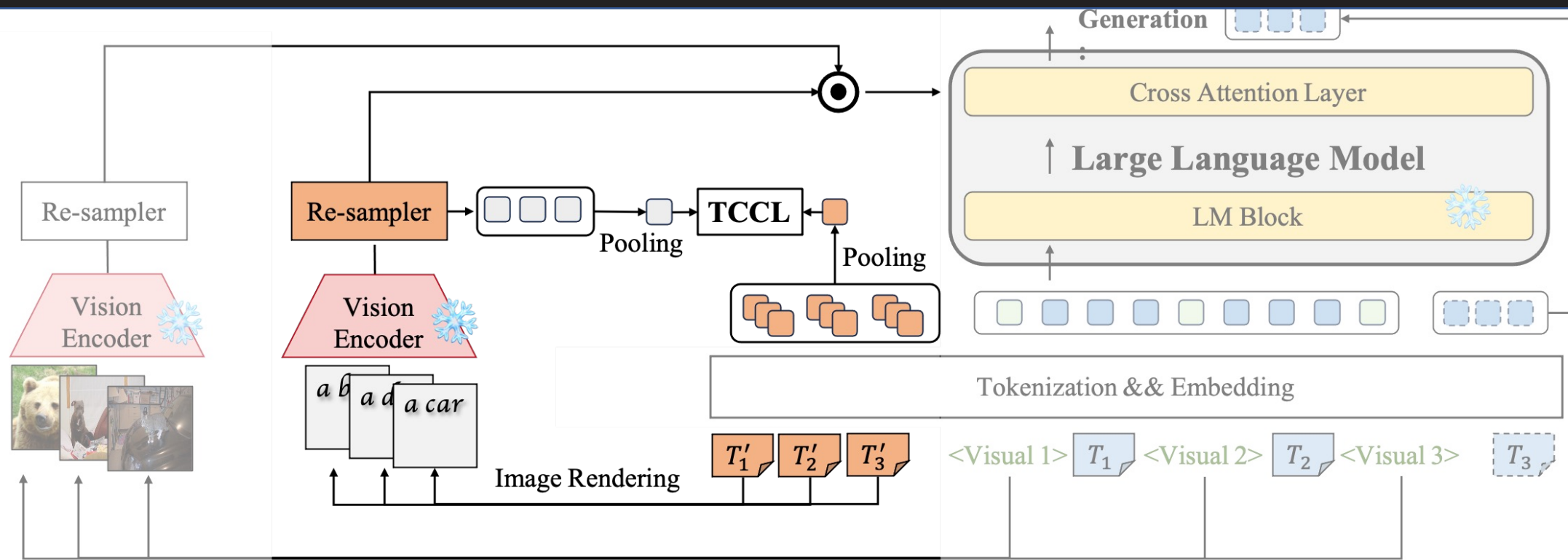
11 11th, 2024

Motivation



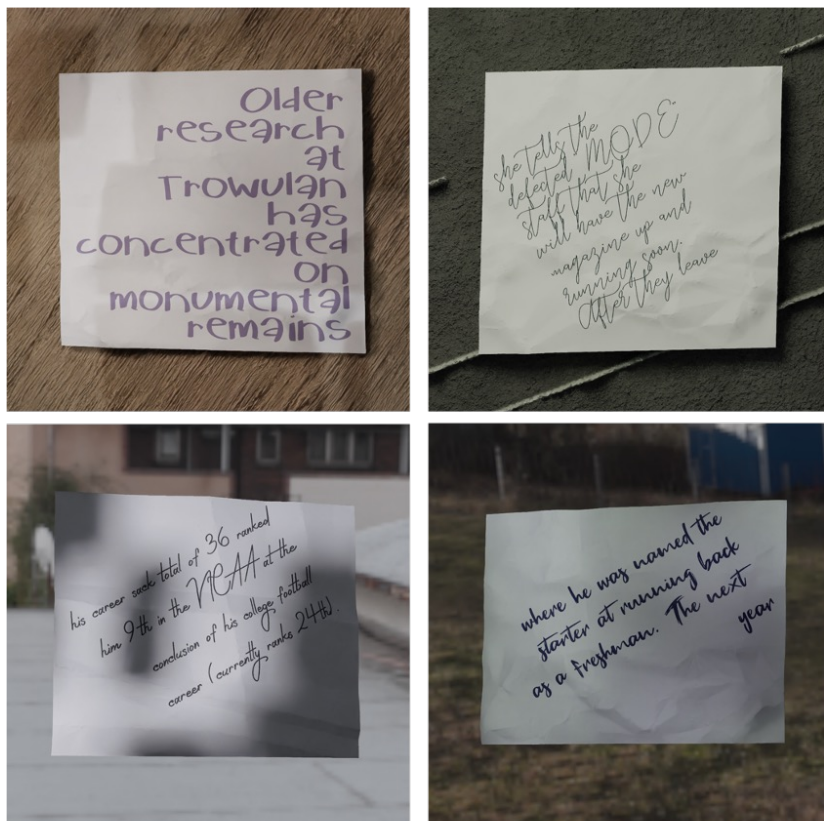
- Existing MLLMs usually exploit **a much lighter visual encoders**, compared to its text decoders
- Visual encoders trained on **paired image-text data also exhibit emergent OCR capabilities**.
- Convert long textual content into images and use the visual encoders to extract textual representations.

Method

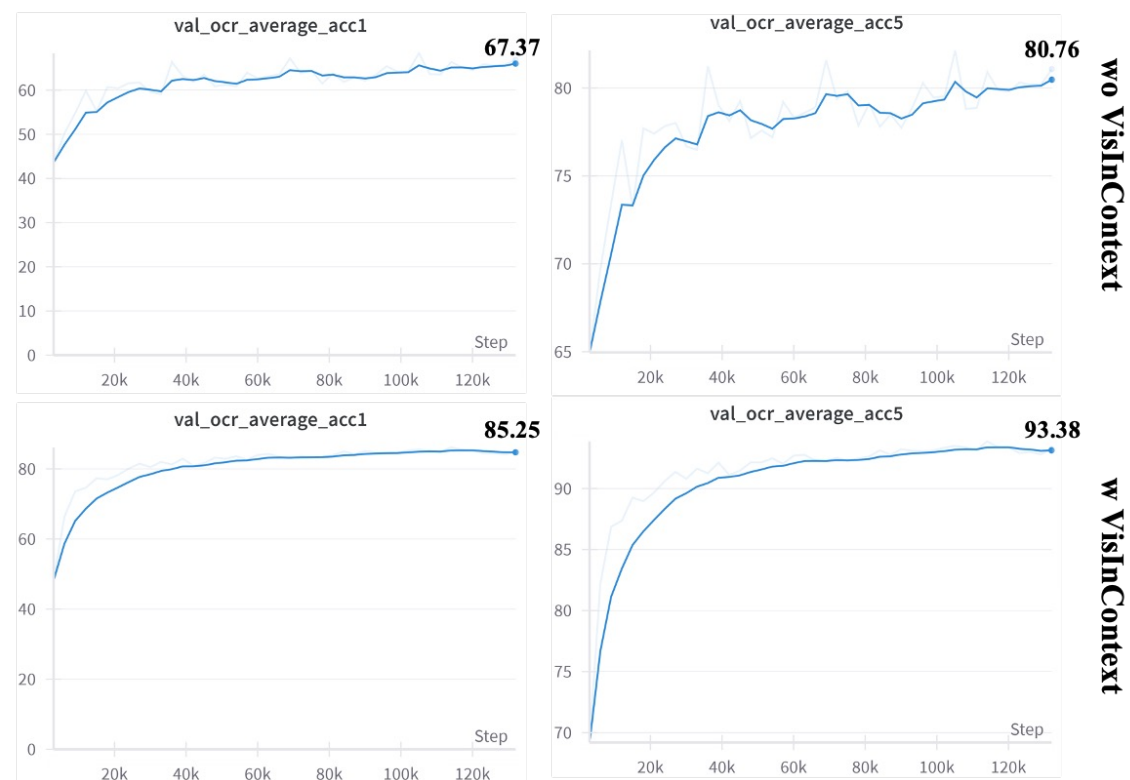


- We convert these omitted text context into visual signals by **rendering them into image**
- VisInContext pipeline builds upon the Flamingo model for **in-context few-shot modeling** (represented in gray)
- VisInContext processes interleaved image-text data by **rendering portions of the in-context text into images**
- **Maintains the Text Token Length** of the model while allowing for a significantly extended *In-context Text Length*.

Experiment



(a). Rendered Text



(b). Token Prediction Accuracy on Validation Set

VisInContext significantly improves the OCR ability of LLM.

VisInContext achieves significantly better results in predicting words in visual images, even when the fonts are difficult to recognize.

Experiment

Genome Structure T_1

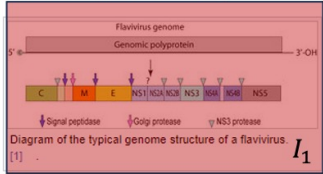


Diagram of the typical genome structure of a flavivirus. [1]

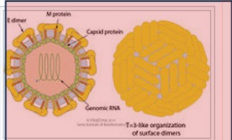
The Zika virus is a positive sense single-stranded RNA molecule 10794 bases long [3] with two non-coding regions flanking regions known as the 5' NCR and the 3' NCR. The open reading frame of the Zika virus reads as follows: 5'-C-prM-E-NS1-NS2A-NS2B-NS3-NS4A-NS4B-NS5-3' and codes for a polyprotein that is subsequently cleaved into capsid (C), precursor membrane (prM), envelope (E), and non-structural proteins (NS). [8] The E protein composes the majority of the virion surface and is involved with aspects of replication such as host cell binding and membrane fusion. [3] NS1, NS3, and NS5 are large, highly-conserved proteins while the NS2A, NS2B, NS4A, and NS4B proteins are smaller, hydrophobic proteins. [8] Located in the 3' NCR are 428 nucleotides that may play a part in translation, RNA packaging, cyclization, genome stabilization, and recognition. [3] The 3' NCR forms a loop structure and the 5' NCR allows translation via a methylated nucleotide cap or a genome-linked protein. [7]

Virion Structure of a Zika virus T_2

The structure of ZIKV follows that of other flaviviruses. It contains a nucleocapsid approximately 25-30 nm in diameter surrounded by a host-membrane derived lipid bilayer that contains envelope proteins E and M. The virion is approximately 40 nm in diameter with surface projections that measure roughly 5-10 nm. [8] The surface proteins are arranged in an icosohedral-like symmetry. [7]

Reproductive Cycle of a Zika virus in a Host Cell

The reproductive cycle of ZIKV follows that of other known flaviviruses. First, the virion attaches to the host cell membrane receptors via the envelope protein which induces virion endocytosis. Next, the virus membrane fuses with the endosomal membrane and the ssRNA genome of the virus is released into the cytoplasm of the host cell. It is then translated into a polyprotein that is subsequently cleaved to form all structural and non-structural proteins. Replication then takes place at intracellular compartments known as cytoplasmic viral factories in the endoplasmic reticulum resulting in a dsRNA genome. The dsRNA genome is then transcribed resulting in additional ssRNA genomes. Assembly then occurs within the endoplasmic reticulum and the new virions are transported to the Golgi apparatus and then excreted into the intracellular space where the new virions can infect new host cells. [7]



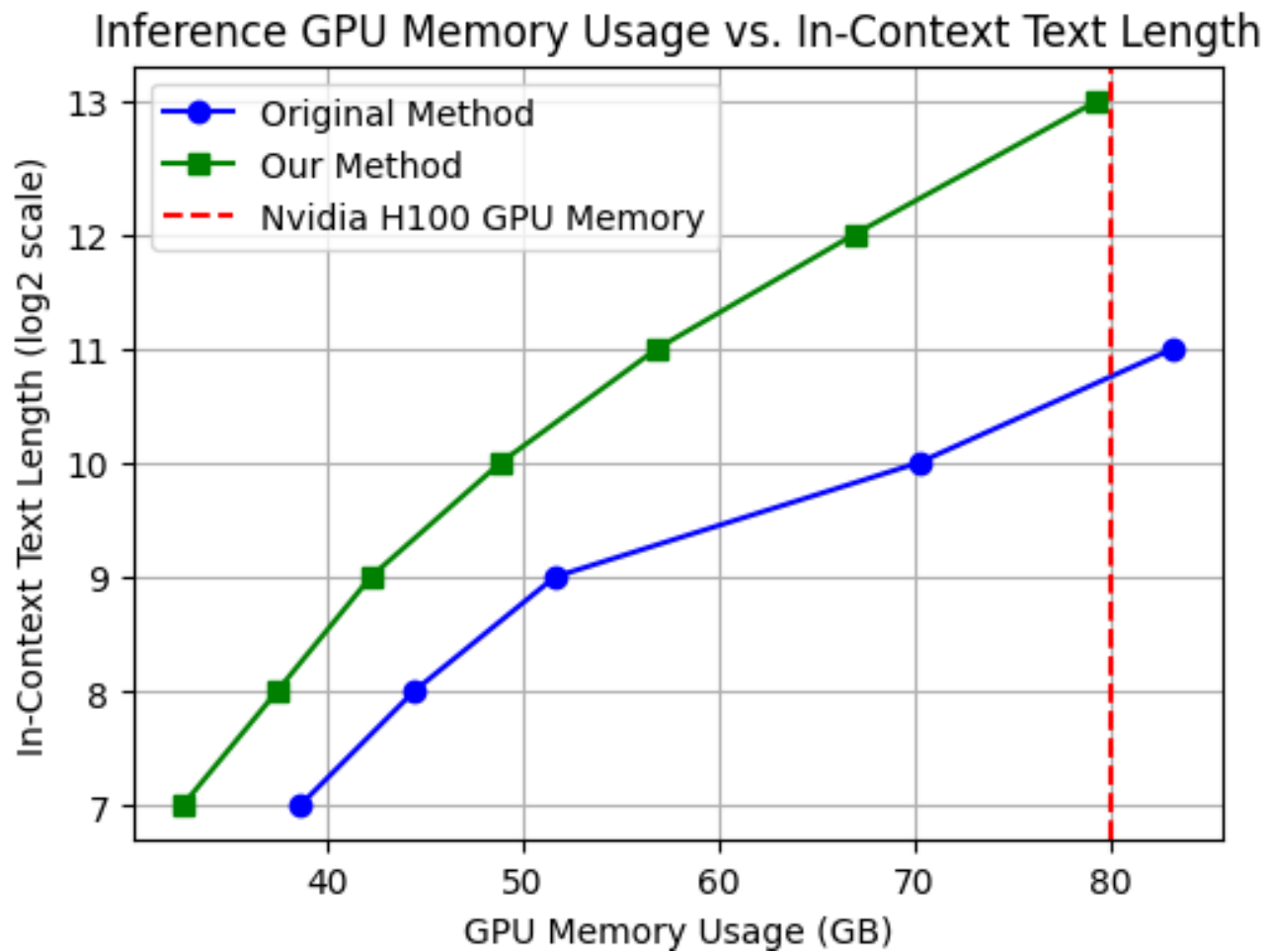
General structure of a flavivirus virion. [2]

Sequential multi-modal retrieval example.

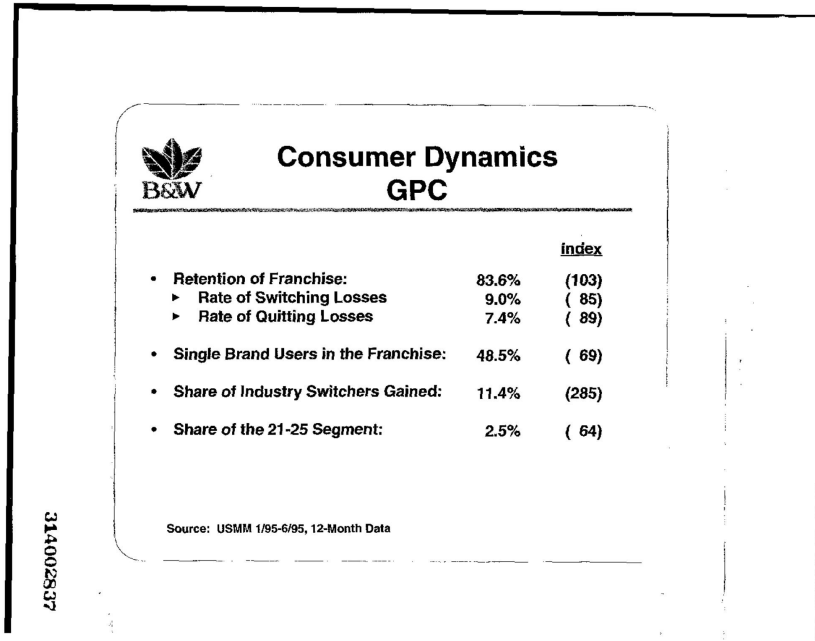
Visual Input	Text Input	Surrounding Text Input	Seq-I	Seq-T
Raw Image	Raw Text	-	16.3	64.8
Raw Image	Raw Text	Raw Text	18.9	67.5
Raw Image	Raw Text	Rendered Text Image	22.7	66.5

The model pretrain with VisInContext significantly improves sequence understanding ability.

Experiment



Experiment



Source: <https://www.industrydocuments.ucsf.edu/docs/rzby0037>

Q: What is the percentage of the share of the 21-25 segment?

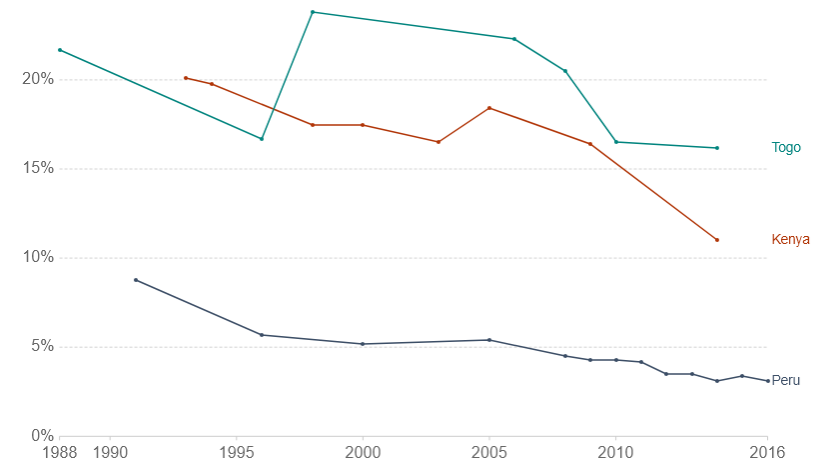
Baseline: 64

+VisInContext : 2.5%

Share of children younger than 5 who are underweight for their age, 1988 to 2016

Our World in Data

Prevalence of underweight children is the percentage of children under age 5 whose weight for age is more than two standard deviations below the median for the international reference population ages 0-59 months.



Source: World Bank

OurWorldInData.org/hunger-and-undernourishment/ • CC BY

Q: Which country is represented by the red line?

Baseline: The red line is Peru

+VisInContext : Kenya

Contribution

- 1. We introduce Visualized In- Context Text Processing (VisInContext), a novel method **that increases in-context text length using visual tokens**.
- 2. We demonstrate that VisInContext is **effective for both training and inference stage** with much lower computational cost.
- 3. With extended text context brought by VisInContext, our model improves the average **in-context few-shot performance** from 55.8% to 57.8% over the competing baseline.
- 4. As a byproduct, our method also **shows great potential in document understanding on popular document QA tasks** and our newly proposed sequential document retrieval task.