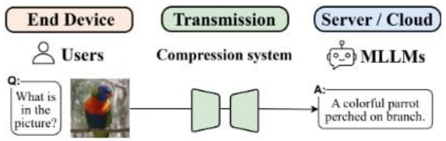
Bridging Compressed Image Latents and Multimodal Large Language Models

Chia-Hao Kao[1,2], Cheng Chien[2], Yu-Jen Tseng[2], Yi-Hsin Chen[2], Alessandro Gnutti[1], Shao-Yuan Lo[3], Wen-Hsiao Peng[2], Riccardo Leonardi[1]

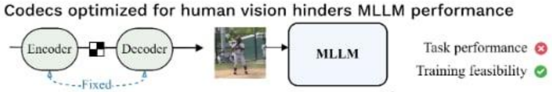
[1]University of Brescia, Italy [2]National Yang Ming Chiao Tung University, Taiwan [3]Honda Research Institute, USA

Motivation

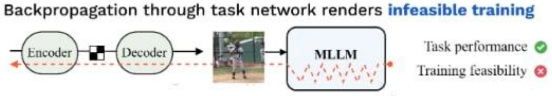
- MLLMs excel at reasoning with text and images but typically require server-side hosting due to their size, necessitating **data transmission**
- **Compression is necessary:**
  w/o compression → 24 bits per pixel (22MB for a 720p image)
  with compression → ≈0.2 bits per pixel (≈0.15MB) ← **120x smaller**



- Codecs optimized for human vision hinders MLLM performance



- Existing coding for machines methods do not consider MLLMs
- Backpropagation through task network renders **infeasible training**



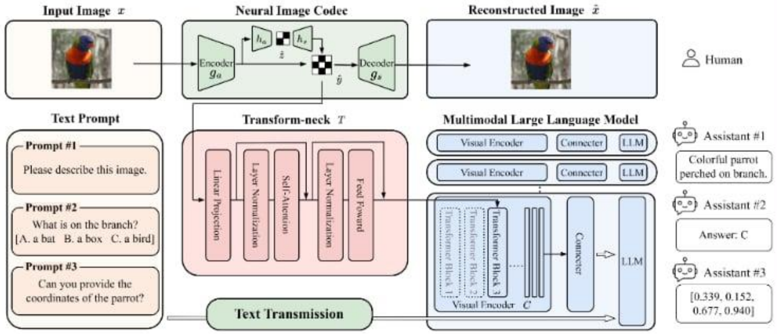
➡ Call for an efficient compression system designed for MLLMs

Contributions

- This marks the **first exploration** of neural image coding for MLLMs
- It adapts image latents directly to MLLMs **saving computational complexity**, while avoiding backpropagating through the heavy MLLM
- It is broadly applicable to a wide range of codecs, tasks, and MLLMs
- It is able to accommodate **various application scenarios** that involve human perception, machine perception, or both

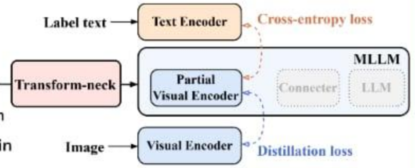
Architecture

- Directly **adapt compressed image latent representations** to connect with MLLMs

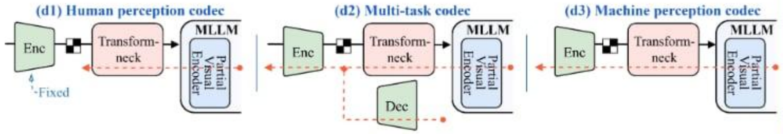


Surrogate Loss

- Leverage **only visual encoder**, avoiding backpropagation through entire MLLM
- **Distillation loss** ensures alignment of the visual features before and after compression
- **CE loss** bridges features with the text domain

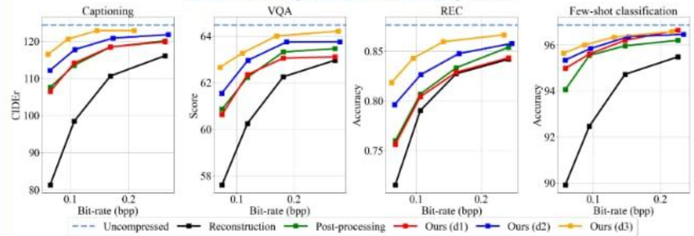


Application Scenarios



Experimental Results

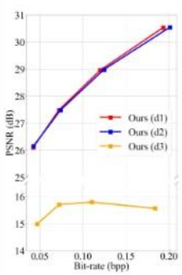
Rate-Accuracy Performance Comparison



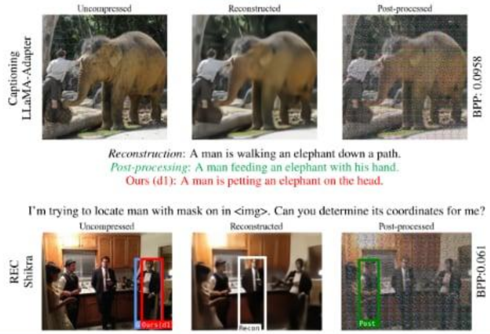
Computational Complexity Comparison

Method	Component	Params (M)		kMAC/pixel	
Ours (d1, d2, or d3)	Transform-neck	13.19		52.795	
Post-processing	Decoder Post-processing network First 2 layers of visual encoder	7.34 31.04 25.78	64.16 (+386%)	112.00 835.72 70.24	1017.96 (+1828%)

Rate-Distortion Comparison



Visualization



Paper, code, project page