Evaluating Efficient Vision-Language Models (VLMs)

LLaVA-1.5 for Descriptive Image Captioning

Daniel Martin Pühringer Patrick Ennemoser Dragana Sunaric

TU Wien

191.021 Introduction to Computational Sustainability

Winter Semester 2025

Goal

- Evaluate LLaVA-1.5 on COCO Captions dataset.
- Compare trade-offs between:
  1. Baseline FP16
  2. Weight-only INT8
  3. Quantized vision tower (INT8)
- Assess:
  - Efficiency (VRAM, latency, throughput, size)
  - Quality (CIDEr score)
  - Energy (CO_2 eq. in kg)

Dataset: COCO 2017 validation

- Dataset: COCO 2017 validation (5,000 images)
- Used first 100 images for experiments

Figure: Example Image: STOP sign

Ground truth captions:

- A stop sign is mounted upside-down on it's post. "
- 'A stop sign that is hanging upside down.'
- 'An upside down stop sign by the road.'
- 'a stop sign put upside down on a metal pole '
- 'A stop sign installed upside down on a street corner'

Quantitative Results

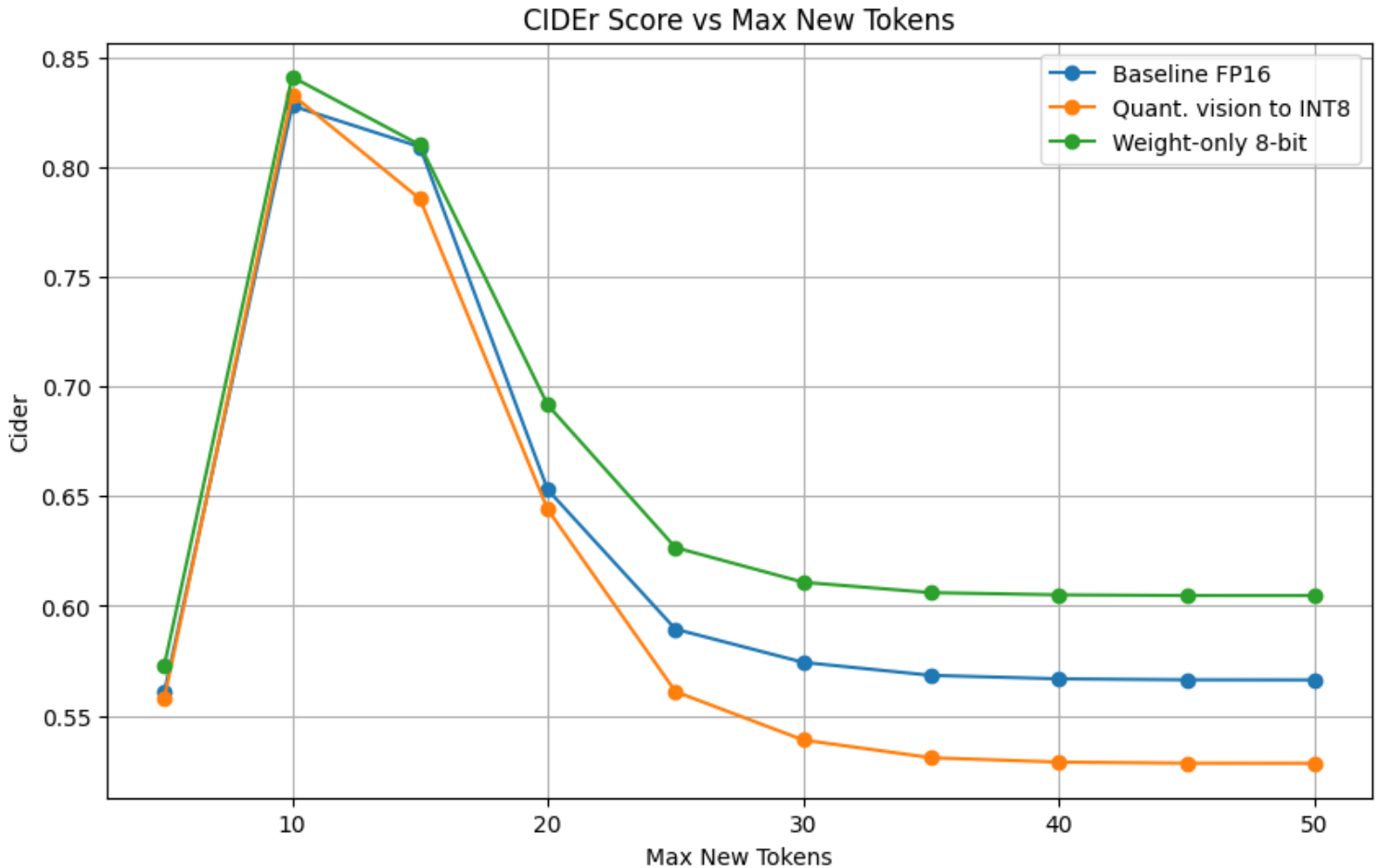

Figure: CIDEr score comparison across models.

Latency and Throughput

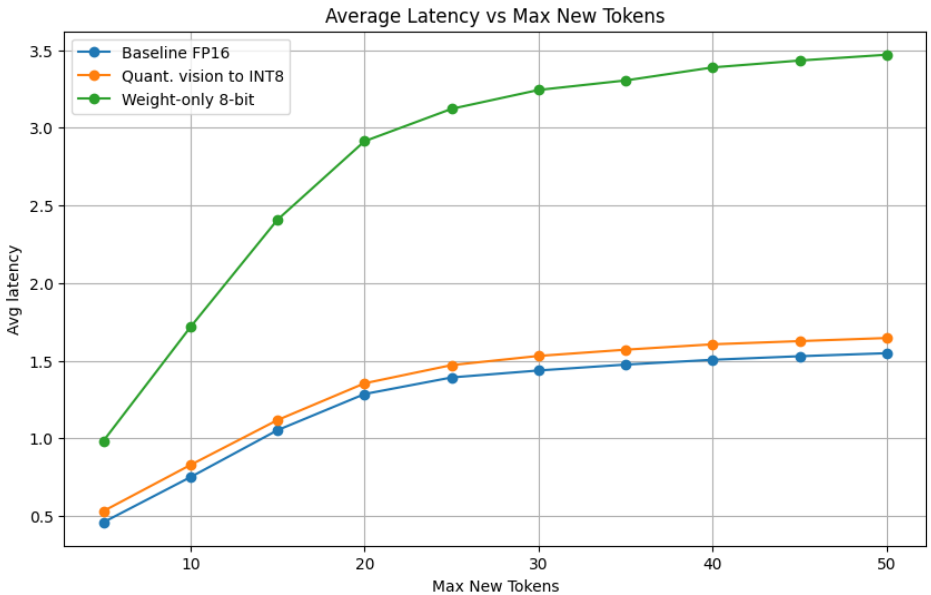

Figure: Latency comparison

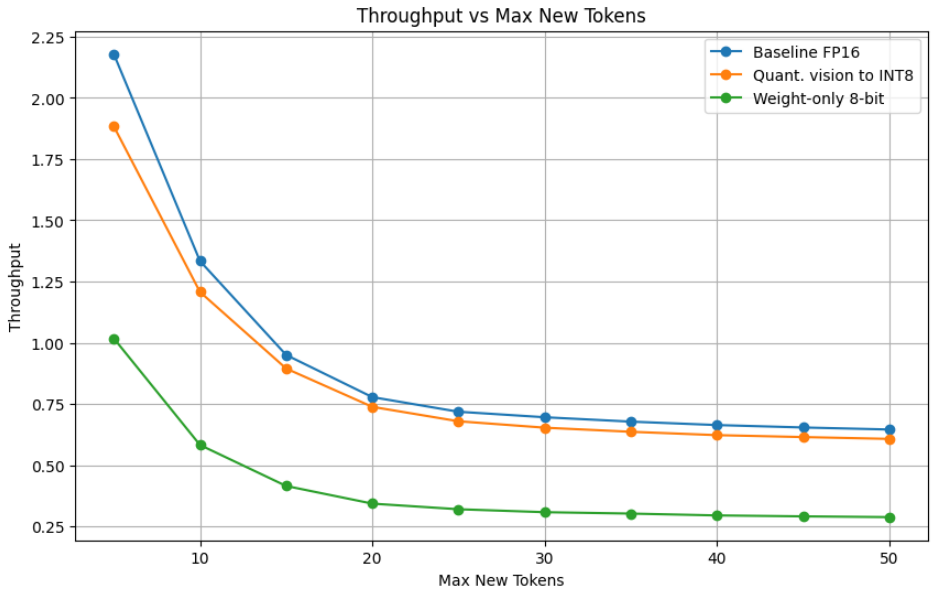

Figure: Throughput comparison

Hardware and Environmental Metrics

Config	VRAM [MiB]	Disk [MiB]	CO_2 eq. [kg]
FP16	13,976.20	55,538.21	0.00610
Weight-only 8-bit	7,538.43	55,663.21	0.01404
Vision Tower INT8	14,846.11	56,539.10	0.00666

Table: Efficiency Metrics

Qualitative Evaluation

Max new tokens	Latency	Generated caption
5	0.434s	A stop sign is ups
10	0.732s	A stop sign is upside down, with the
15	1.038s	A stop sign is upside down, with the word "STOP"
20	1.276s	A stop sign is upside down, with the word "STOP" written backwards.
25	1.276s	A stop sign is upside down, with the word "STOP" written backwards.
30	1.276s	A stop sign is upside down, with the word "STOP" written backwards.
35	1.277s	A stop sign is upside down, with the word "STOP" written backwards.
40	1.278s	A stop sign is upside down, with the word "STOP" written backwards.
45	1.278s	A stop sign is upside down, with the word "STOP" written backwards.
50	1.278s	A stop sign is upside down, with the word "STOP" written backwards.

Table: Generated captions of the FP16 (Baseline) configuration with different token sizes and latency measured in seconds

Figure: Example Image: STOP sign

Experiment configuration	Generated caption for image 6 with 30 max token
FP16 (Baseline)	A group of teddy bears are sitting on a bed, with one teddy bear sitting on top of another.
Weight-only INT8	A group of teddy bears are piled on top of each other, with one teddy bear in the center.
Quant Vision Tower	A group of teddy bears are sitting on a bed, with one teddy bear sitting on top of another teddy bear.

Table: Generated captions for teddy bears image of all three different experiment configurations under a max token size of 30

Figure: Example Image: Teddy Bears

Discussion

- **Performance:** Weight-only 8-bit model shows higher latency and lower throughput due to INT8$\rightarrow$FP16 upcasting;
- **Efficiency:** Weight-only 8-bit model doubles CO_2 emissions.
- **Limitations:** Results based on T4/L4 GPUs under limited settings;

Conclusion

- Quantizing only the vision tower preserves quality and efficiency.
- Weight-only INT8 shows higher latency due to upcasting overhead.
- Further work: combine mixed-precision quantization and distillation.

Questions?