

Image Captioning on a Laptop: Comparing the Quality and Efficiency of Local AI Models

Kshitij Khandagale
Independent Student Research

October 2026

Abstract

This study compares four AI models tasked with generating image captions. The aim was to find out how well they write captions and how much time and GPU memory they use. Each model described the same 120 images, giving 480 captions in total. The models were asked to write eight short phrases separated by commas. Two AI judge models checked each caption twice for accuracy, coverage, and clarity. Qwen3.8 27B received the highest average quality score at 4.89 out of 5. Gemma 4 12B scored 4.79 and generated captions about 2.5 times faster. Qwen3-VL 2B was fastest, but it did not follow the required format. The results suggest that Gemma 4 offered a useful balance between quality and speed in this test.

Keywords: image captioning, vision-language models, local AI, data labeling, large language models, open-source models

1 Introduction

An image label can contain more detail than a word such as dog or car. For example, a caption might describe a person holding an apple, wearing a blue jacket, and standing in a garden. Labels like these can help describe the contents of an image dataset.

Writing detailed captions by hand takes time. A vision-language model, or VLM, is an AI model that can look at an image and respond to a written instruction. Running these models locally may help speed up the creation of image-caption datasets.

However, a useful model must do more than name the main object. Its captions should be accurate and follow the requested format, and the model should run within the available hardware. This study asks: *How do the tested local vision-language models compare in caption quality, adherence to the required output format, and computational efficiency?*

Code and data: <https://github.com/kprk369/vlm-caption-benchmark>

2 Related Work

Image captioning has been studied for many years. Show and Tell used a neural network to turn images into sentences [4]. Later systems such as LLaVA showed how models could follow written instructions about images [3].

Evaluating captions is difficult because different wording can describe the same image correctly. CLIPScore showed a way to compare a caption directly with an image without needing a reference sentence [2]. Another approach is to ask an AI model to rate the output. In experiments on text-based assistants, Zheng et al. reported agreement with human preferences, but also biases related to answer order and length [5]. Those findings do not establish the accuracy of this study’s image-caption judges.

3 Method

3.1 Images and models

The dataset contained 120 images from Unsplash Lite. There were 20 images in each of six groups: objects and technology, people and actions, indoor scenes, outdoor scenes, animals, and complex or text-heavy scenes. The images were checked for duplicates and damaged files.

Four caption models were tested: Qwen3.8 27B Q4_K_M, Gemma 4 12B Q8_0, Qwen3-VL 2B Q8_0, and Gemma 3 4B Q8_0. Quantization, shown by the Q4 and Q8 labels, reduces how much memory a model needs. Because the models used different sizes and quantizations, this experiment compares these particular versions rather than model size alone.

All caption generation used Ollama 0.40.3 on the same Windows laptop. It had an RTX 5090 Laptop GPU with about 24 GB of video memory, an Intel Core Ultra 9 275HX processor, and about 64 GB of RAM.

3.2 Generating the captions

Each model described all 120 images once. The same prompt asked for exactly eight unique, factual phrases on one line, separated by commas. Each phrase had to contain fewer than ten words. The model was asked to describe visible objects, actions, colors, relationships, the setting, and readable text without guessing.

The caption settings were the same for every model: seed 42, generation temperature 0, a 4,096-token context, and a maximum output of 256 tokens. This was a limit, not a required caption length. A token is a small piece of text processed by the model. Captions were kept as generated, including any mistakes or formatting problems. The full caption prompt is reproduced in Appendix A.

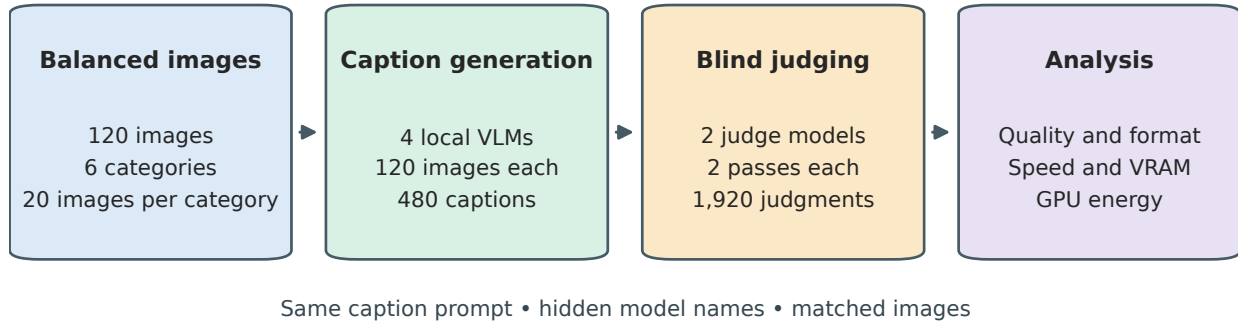


Figure 1: The main steps of the experiment.

3.3 Checking the captions

Two local models acted as judges: Qwen3.8 27B Q4 and Qwen3.8-Flash-Next IQ3_XXS. Each judge checked every caption twice. The four passes used seeds 101, 202, 303, and 404, with generation temperature 0.2. This produced 1,920 evaluations.

The judges saw the image and caption, but were not told which model wrote it. They rated three features from 1 to 5: **accuracy**, meaning whether the details were correct; **coverage**, meaning whether the important parts were included; and **clarity**, meaning whether the text was easy to understand and avoided repetition.

The quality score was the average of these three ratings across all four passes. The judges also decided whether a caption was usable without editing. The full judge rubric and the way it was supplied to each judge are given in Appendix A.

A separate **format check** required nonempty text on one line with exactly eight nonempty comma-separated phrases and no duplicate phrases after ignoring case and extra whitespace. It did not check every prompt instruction: phrase length, comma-space style, other punctuation, and repeated meanings were not tested. A caption counted as **ready to use** when at least three passes marked it usable and it passed this basic format check. This definition does not guarantee complete prompt compliance or human-verified accuracy.

The analysis also checked how much scores changed between passes. Uncertainty around the model averages was estimated by resampling the 120 images 5,000 times. The analysis code and detailed statistical results are included with the project.

3.4 Measuring computing resources

During each caption request, the benchmark recorded processing time, GPU memory, power, temperature, CPU use, and RAM use when available. GPU energy was estimated from sampled power readings during the request.

The estimates for 100,000 images were calculated separately for each model from its 120 caption requests. Let $\bar{t}$ be the average processing time in seconds per image and $\bar{e}$ be the average estimated

GPU energy in watt-hours (Wh) per image. The projected processing time was

$$T_{100\,000} = \frac{100\,000 \times \bar{t}}{3\,600} \text{ hours.} \quad (1)$$

The projected GPU energy was

$$E_{100\,000} = \frac{100\,000 \times \bar{e}}{1\,000} \text{ kWh.} \quad (2)$$

The divisors convert seconds to hours and watt-hours to kilowatt-hours. Calculations used the unrounded averages. These projections assume images are processed one at a time at the same average rate. The energy estimate covers the GPU during requests, so it does not represent the electricity use of the whole laptop.

The illustrative GPU electricity cost was calculated as $C_{100\,000} = E_{100\,000} \times r$, using $r \approx 0.08437$ USD per kWh. This retains the original assumption of INR 8/kWh, converted using the European Central Bank’s reference rates for 22 October 2026: EUR 1 equaled USD 1.1205 and INR 106.2490 [1]. Calculations used the unrounded conversion. The electricity price is illustrative, not a measured local tariff.

4 Results

4.1 Caption quality and format

Qwen3.8 had the highest average quality score at 4.89 out of 5. Gemma 4 followed at 4.79. Qwen3-VL and Gemma 3 scored 4.31 and 4.27. The difference between the two smaller models was too uncertain to establish a clear quality advantage.

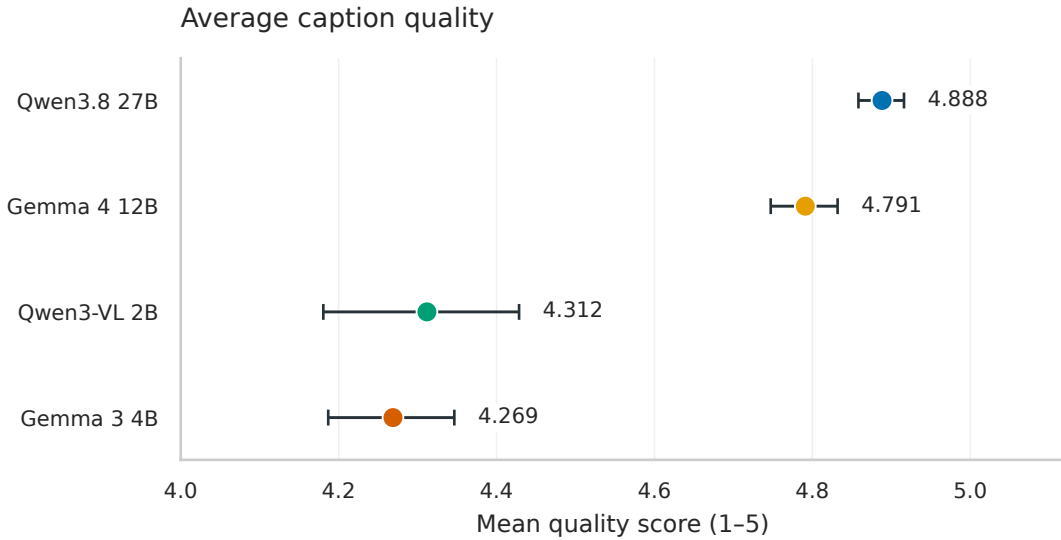


Figure 2: Average caption quality. The lines show 95% uncertainty intervals around each average.

Table 1: Caption quality and readiness for use. Each model produced 120 captions.

Model	Quality / 5	Format check	Ready to use
Qwen3.8 27B	4.89	99.2%	98.3%
Gemma 4 12B	4.79	100.0%	97.5%
Qwen3-VL 2B	4.31	0.0%	0.0%
Gemma 3 4B	4.27	91.7%	65.8%

Qwen3.8 and Gemma 4 produced captions that usually passed both the content and format checks. Gemma 3 had a lower ready-to-use rate of 65.8%.

Qwen3-VL shows why format needed its own check. At least three judging passes marked 85.8% of its captions usable, but none passed the basic format check. Inspection of all 120 saved captions found that none contained commas. There were 80 single-line outputs and 40 eight-line lists. Without commas, the checker counted each output as one phrase rather than eight. The punctuation was also absent from the original saved model responses, so it was not removed during CSV export.

Eight Qwen3-VL outputs reached the 256-token limit. The other 112 ended with the recorded completion reason `stop`, so truncation alone cannot explain the 0% result. This is a formatting failure in the tested configuration, not evidence that every description was factually wrong. The records cannot isolate whether the underlying cause was the model, quantization, or runtime behavior. No captions or original scores were changed during this audit.

4.2 Actual caption examples

The following are unedited outputs for the same image, `image_005.jpg`, shown in Figure 3. These examples were chosen to show differences in output structure.



Figure 3: Image 005, used for all four caption examples below.

Qwen3.8 27B.

Man walking on sandy cave floor, dark layered rock walls, small opening revealing bright sky, white foam in shallow water, scattered gray and brown rocks, black t-shirt with small white logo, tan pants with red patch, man looking downward

Gemma 4 12B.

a man walking, a large sea cave, layered rock walls, sandy beach floor, ocean water waves, dark grey rocks, a small opening, black t-shirt

Qwen3-VL 2B.

man walking on sandy beach rocks cave entrance water waves dark rock layers light on sand person in black shirt brown pants white rocks small stones dark cave entrance

Gemma 3 4B.

Dark cave walls surround a man, sandy beach extends forward, man wears a black shirt, he walks on wet sand, rocks line the cave floor, a small opening illuminates the space, muted brown and gray tones dominate the scene

For this image, Qwen3.8 and Gemma 4 passed the format check. Qwen3-VL used no commas, while Gemma 3 produced seven comma-separated phrases rather than eight. These examples show why readable content and format-check results were reported separately.

In this example, Gemma 4 did not mention that the man was looking downward, while Qwen3.8 included this detail. This illustrates a difference in descriptive coverage. Across the benchmark, Qwen3.8 received a higher average quality score from the AI judges but used more GPU memory.

4.3 Speed, GPU memory, and electricity cost

Qwen3-VL was fastest at 0.78 seconds per image and used the least GPU memory. Qwen3.8 was slowest at 3.77 seconds per image and used the most memory. Gemma 4 generated captions about 2.5 times faster than Qwen3.8 while keeping a similar ready-to-use rate.

Table 2: Caption-generation time, GPU memory use, and projected GPU energy and electricity cost.

Model	Time/image (seconds)	Peak VRAM (GiB)	Time/100k (hours)	GPU energy/100k (kWh)	GPU cost/100k (USD)
Qwen3.8 27B	3.77	16.8	104.8	9.20	0.78
Gemma 4 12B	1.48	13.1	41.2	1.64	0.14
Qwen3-VL 2B	0.78	4.0	21.7	0.61	0.05
Gemma 3 4B	0.99	5.4	27.5	0.95	0.08

Costs assume approximately USD 0.08437/kWh, equivalent to INR 8/kWh at the stated exchange rate. They cover caption-generation GPU energy only, excluding judging, the rest of the laptop, hardware costs, and human review.

Projected GPU energy for 100,000 images was 9.20 kWh for Qwen3.8, 1.64 kWh for Gemma 4, 0.61 kWh for Qwen3-VL, and 0.95 kWh for Gemma 3. These are estimates from the small test, rather than measurements from processing 100,000 images. They exclude the rest of the laptop and cannot be used as a full financial cost comparison.

5 Discussion and Limitations

Qwen3.8 was the strongest model when judged caption quality was the main goal. Gemma 4 offered a useful balance because its score was close to Qwen3.8’s while it was faster and used less GPU memory. This is a practical interpretation of the results, rather than a separate measured score.

The smaller models required less memory and time, but their outputs needed more care. Gemma 3 had lower quality and usability. Qwen3-VL produced useful content in many cases, but its formatting problems would require another processing or editing step.

The main limitation is the use of AI judges. Both judges belong to the Qwen3.8 family, and Qwen3.8 also generated some of the captions. This may favor related outputs. Repeating evaluations helps measure consistency, but it does not prove that the judgments are correct.

The judges agreed on the usable decision for about 90% of captions when compared with each other. Their exact agreement on the three quality ratings was lower, ranging from about 60% to 78%. The scores should therefore be treated as AI assessments rather than verified human accuracy.

One additional field asked the Qwen3.8 judge for confidence from 0 to 1. It returned values outside that range in 957 of 960 evaluations. Those confidence values were excluded. The main quality ratings were retained because they passed their own range checks.

The sample was limited to 120 images and one prompt on one laptop. Other images, settings, or hardware could give different results. No matched human labeling test was completed, so this study cannot show that AI labeling is better or cheaper than human labeling.

The project includes the prompts, code, settings, generated captions, and evaluation records so that the analysis can be repeated. The full licensed Unsplash image dataset is kept separately and is not redistributed with the repository. The example image shown in this local manuscript is also excluded from the repository.

6 Conclusion

This study compared four local AI models for detailed image captioning. Qwen3.8 gave the highest judged quality, while Gemma 4 offered a useful balance of quality, speed, and GPU memory use. The results also showed that a good description can still fail the required output format.

Local models may help create image labels, but these results need human validation. A useful next step would be to have people check a smaller sample and record the time needed to review and correct the captions.

AI Assistance Disclosure

Generative AI tools were used to assist with code development, experimental implementation, data analysis, and manuscript preparation. The author directed the research process, supervised the experiments, reviewed the results, and evaluated the findings and limitations. AI-generated outputs were reviewed as part of the research workflow. The author assumes responsibility for the methodology, reported results, interpretations, and final manuscript.

References

- [1] European Central Bank. Euro foreign exchange reference rates. https://www.ecb.europa.eu/stats/policy_and_exchange_rates/euro_reference_exchange_rates/html/index.en.html, 2026. Reference rates for 22 October 2026: EUR 1 = USD 1.1205 and INR 106.2490. Accessed 23 October 2026.
- [2] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021. <https://doi.org/10.18653/v1/2021.emnlp-main.595>.
- [3] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. <https://doi.org/10.52202/075280-1516>.
- [4] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and Tell: A Neural Image Caption Generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015. Author preprint: <https://arxiv.org/abs/1411.4555>.
- [5] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. <https://doi.org/10.52202/075280-2020>.

A Prompts

The text below reproduces the frozen prompts used in the benchmark.

A.1 Caption-generation prompt

All four caption models received this text as a user message with the image attached. No model-specific wording was added.

Describe only what is visibly present in the image.

Output exactly one line containing exactly 8 unique factual phrases. Separate the phrases using a comma followed by a space. Do not use commas inside a phrase. Do not use sentences, periods, semicolons, bullets, numbering, or line breaks.

Keep every phrase under 10 words. Cover the main objects, actions, colors, spatial relationships, setting, and clearly readable text. Do not repeat information. Do not list absent things. Do not guess identities, locations, brands, or unreadable text.

Output only the 8 comma-separated phrases.

A.2 Judge prompt

Both judge models used the same rubric below in both passes. The seeds differed between passes, but the rubric did not change. The rubric was supplied in a user message to Qwen3.8 through Ollama and in a system message to Flash-Next through Strata.

You are an impartial evaluator of image captions for a research benchmark.

Evaluate the candidate caption using only the supplied image and caption. Do not infer which model wrote it. The caption generator was explicitly instructed to return one line containing exactly 8 unique comma-separated factual phrases. Therefore, comma-separated phrases are the required format: never lower Clarity merely because the caption is a list rather than sentences.

A claim is accurate only when it is visibly supported. Treat invented objects, actions, colors, spatial relationships, brands, identities, locations, and unreadable text as errors. Do not reward verbosity.

Accuracy (1-5)

- 1 = major contradictions or extensive hallucination
- 2 = several important factual errors
- 3 = mostly correct, but at least one meaningful error or unsupported detail
- 4 = accurate with at most one minor factual issue
- 5 = every substantive claim is visibly supported

Coverage (1-5)

- 1 = misses the central subject or action
- 2 = captures little beyond the central subject
- 3 = covers the central content but misses several salient elements
- 4 = covers nearly all important visible elements

5 = covers all important elements reasonably expected within 8 phrases

Clarity (1-5)

- 1 = confusing, severely repetitive, contradictory, or malformed
- 2 = difficult to read or substantially repetitive
- 3 = understandable but noticeably awkward or repetitive
- 4 = clear with only a minor wording or repetition issue
- 5 = concise, readable, non-repetitive, and internally consistent

Usable means the caption could serve as a factual training label without editing. Mark it unusable for a substantive hallucination or contradiction, an important wrong relationship/action/readable text claim, or severe repetition/malformed output. Do not mark it unusable solely for a small color nuance, subjective style word, or omission of a non-central detail.

Return short fields. Identify only the single most important error, omission, and clarity issue. Use an empty string when none exists. Keep the rationale to one concise sentence.

B Supplementary Results

These figures provide additional detail on the quality ratings and format-check results for the 120-image test.

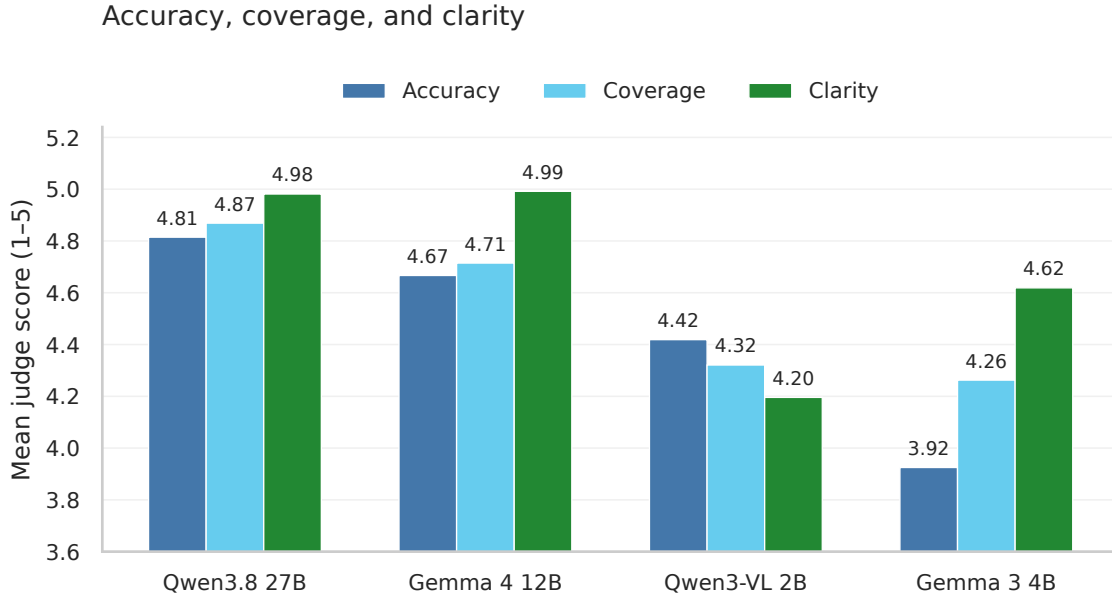


Figure S1: Mean accuracy, coverage, and clarity scores from the four AI judging passes. Each model captioned the same 120 images; ratings used a 1–5 scale.

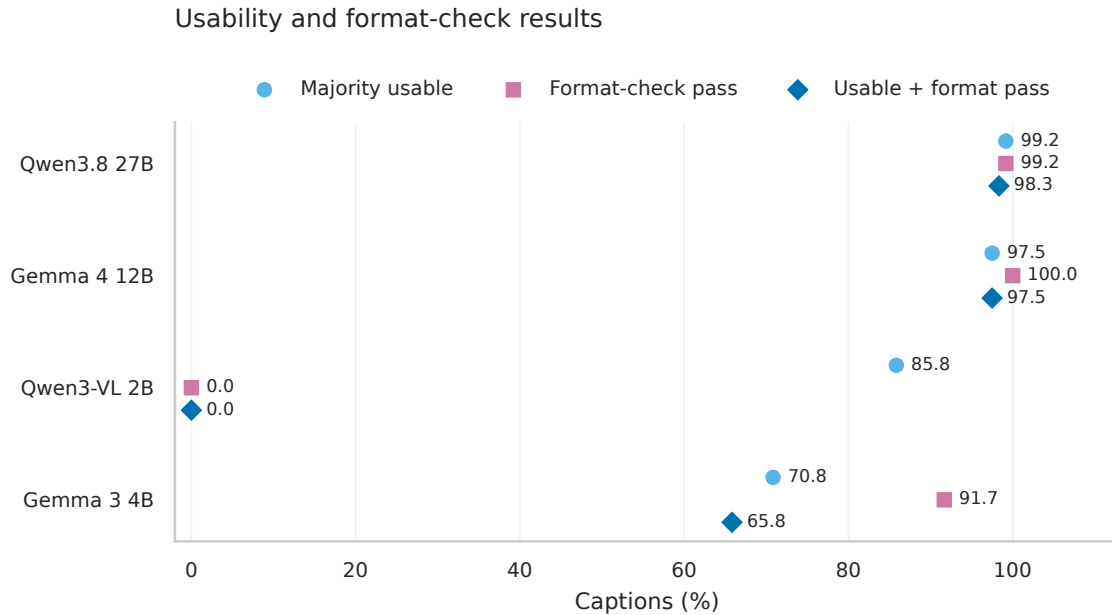


Figure S2: Percentages of captions marked usable by at least three judging passes, passing the basic format check, and meeting both conditions. Passing this check does not guarantee compliance with every prompt instruction.