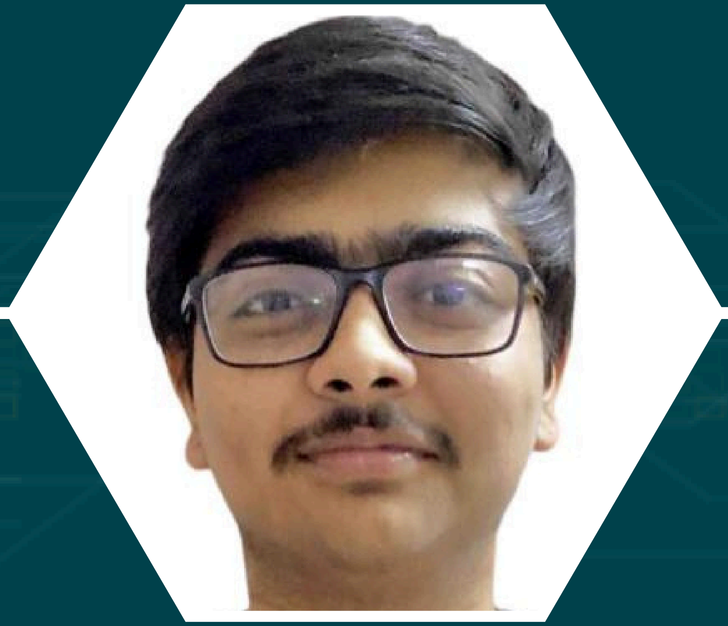
Amazon ML Challenge Finale
Team DBkaScam

Arnav Goel
IIIT-Delhi

Medha Hira
IIIT-Delhi



Mihir Agarwal
IIT-Gandhinagar

A.S. Poornash
IIT-Patna

Dataset Details

Dataset Download and Processing Method:

- We accessed and utilized the data by employing sharding.
- The test set was divided into 13 chunks, each approximately 10,000 samples in size.

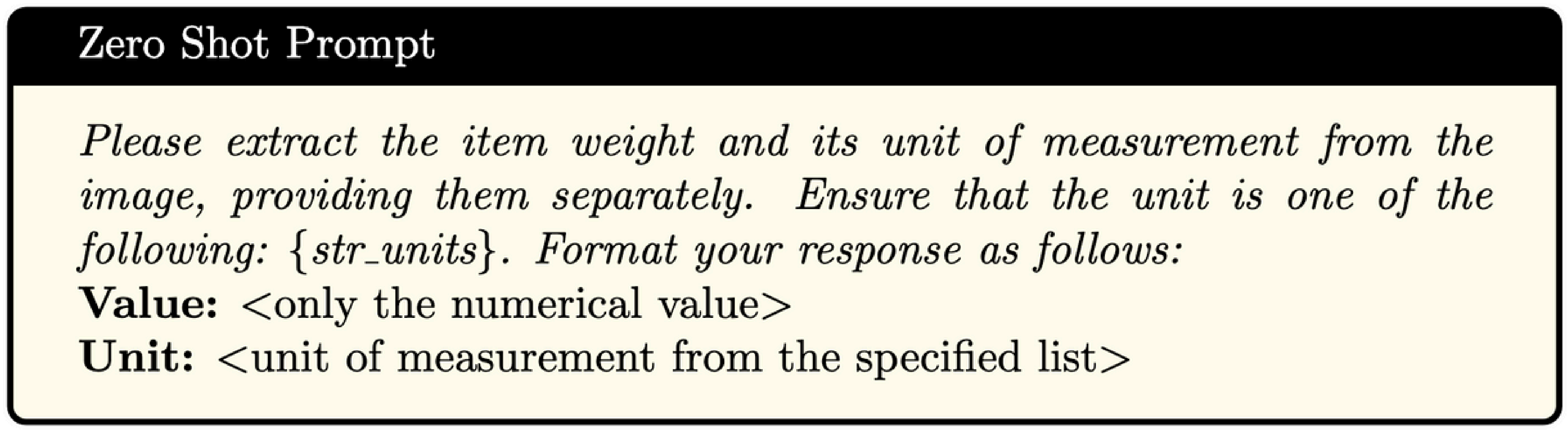
Initial Approaches and Drawbacks

Approach	Drawback
Multi-task Regression+Classification	Suboptimal F1 metric with ±2% deviation limit, leading to poor results with regression.
Tesseract-based OCR + Open-Source LLM (like LLAMA-3.1)	Struggled with poor image feature extraction, compounding errors. Max F-1 score on train-set split < 0.55

Vision Language Model

Name	Language Model	Vision Model	Parameters
MiniCPM-V-2.6	Qwen2-7B	SigLIP-400M	8B
Qwen2-VL-7B	Qwen2-7B	ViT-600M	7B
InternVL2-8B	InternLM2.5-7B	InternViT-300M	8B

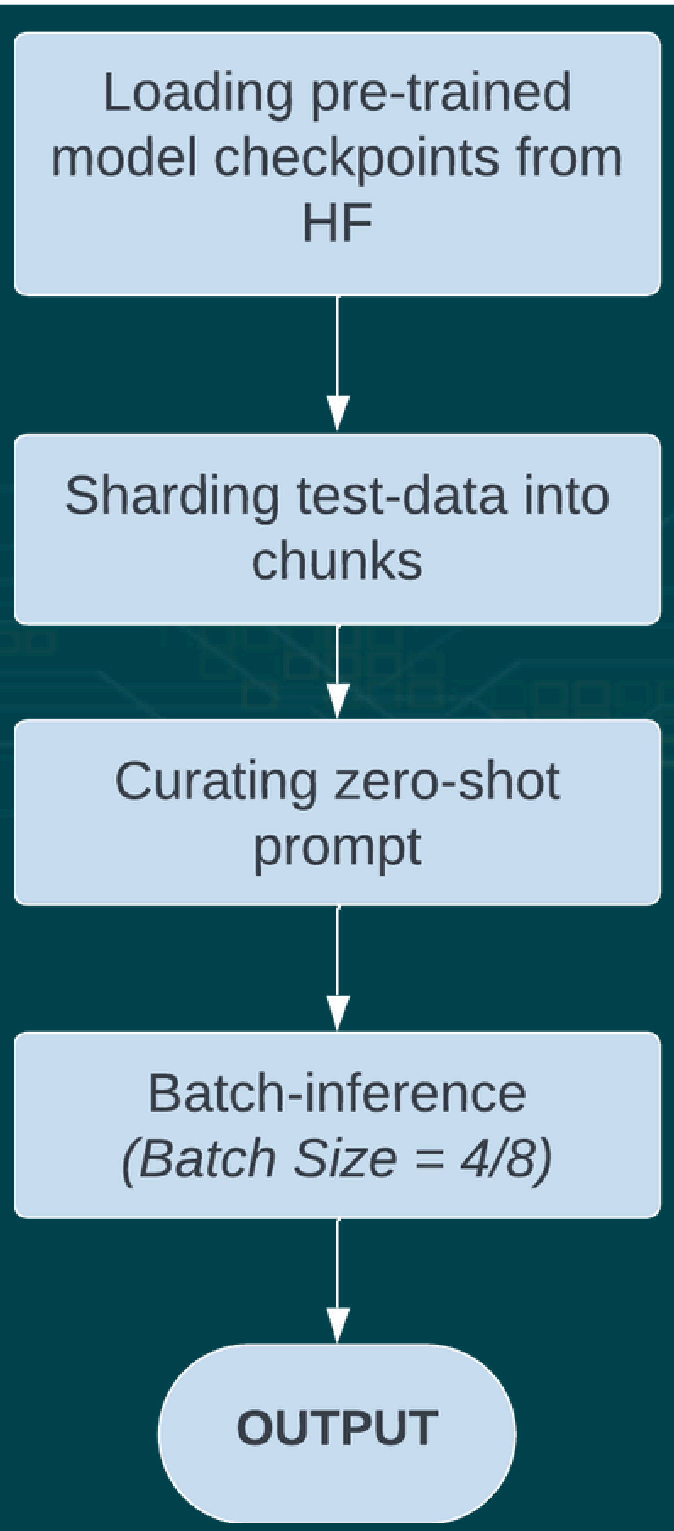
VLMs selected based on performance on OpenVLM Leaderboard (Sorted by OCRBench Performance)



Zero Shot Prompt

Please extract the item weight and its unit of measurement from the image, providing them separately. Ensure that the unit is one of the following: {str_units}. Format your response as follows:
Value: <only the numerical value>
Unit: <unit of measurement from the specified list>

Inference Pipeline



Loading pre-trained model checkpoints from HF

Sharding test-data into chunks

Curating zero-shot prompt

Batch-inference *(Batch Size = 4/8)*

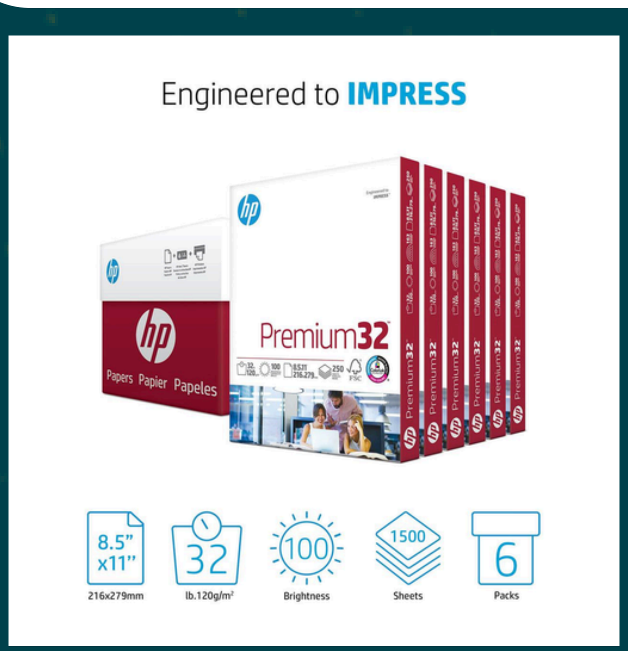
OUTPUT

Few-Shot Learning

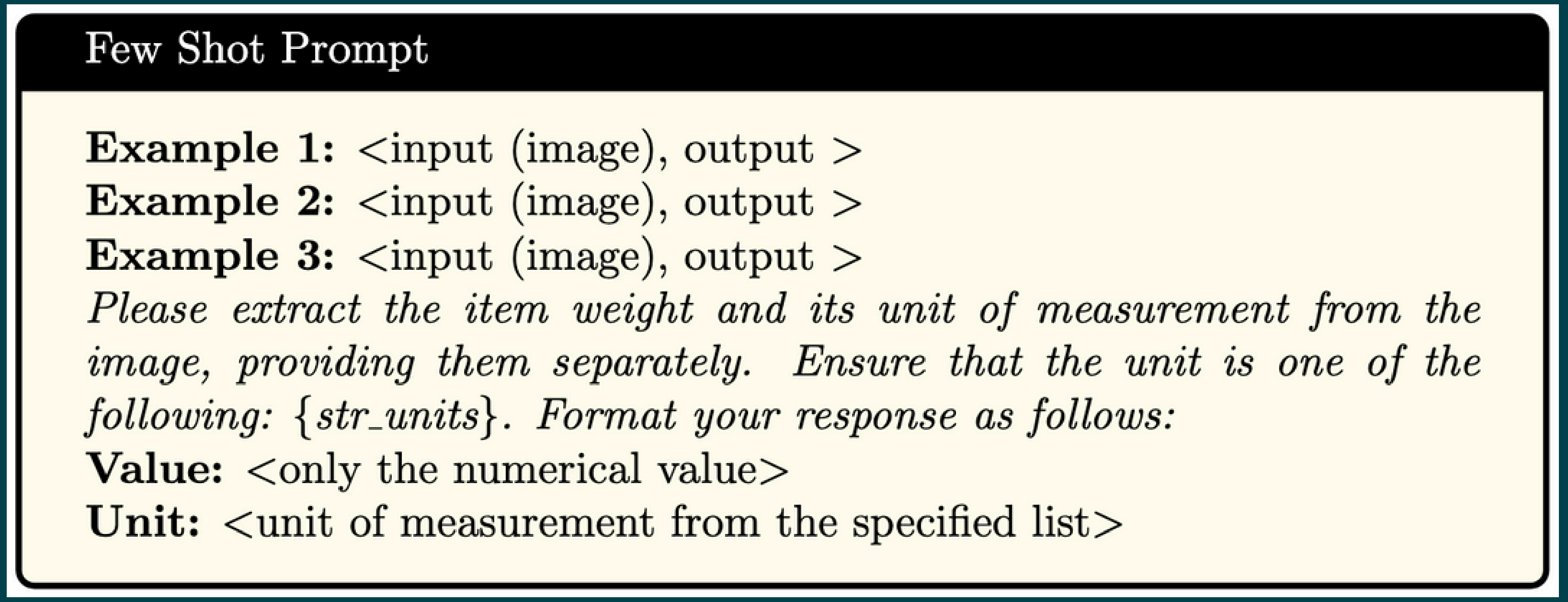
Total Digest contains natural, plant-derived digestive enzymes that:
Break down the fillers in pet food
Support normal digestion
Help pets get more vitamins & nutrients from their food
Support energy levels & immune function
Thomas Pet
TOTAL DIGEST

Entity: Weight

Entity: Height



Entity: Weight



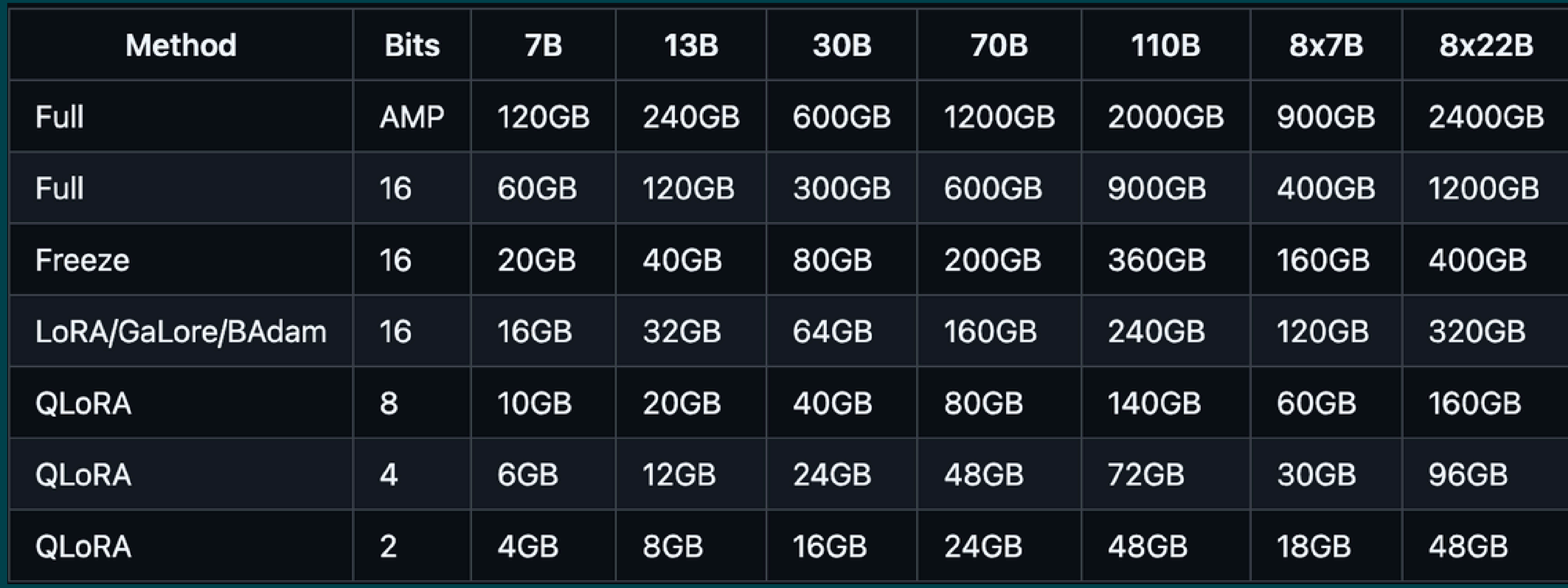
Few Shot Prompt

Example 1: <input (image), output >
Example 2: <input (image), output >
Example 3: <input (image), output >
Please extract the item weight and its unit of measurement from the image, providing them separately. Ensure that the unit is one of the following: {str_units}. Format your response as follows:
Value: <only the numerical value>
Unit: <unit of measurement from the specified list>

Supervised Fine-Tuning

- We utilised LLaMa-Factory (Zheng et al., 2024) for performing parameter-efficient SFT on Qwen2-VL-7B.
  - Q-LoRA : 8-bit quantisation
  - LoRA : 16-bit quantisation
- Source for Table: LLaMa-Factory GitHub
- Fine-Tuned on 150000 samples with a batch size of 16 for 1 epoch.

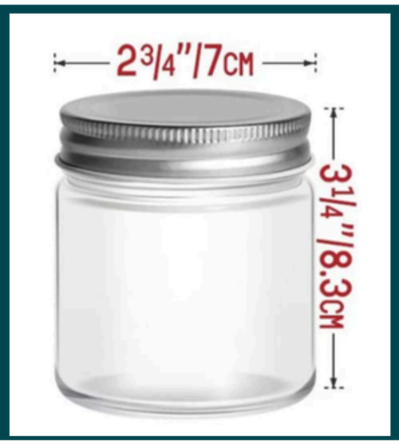
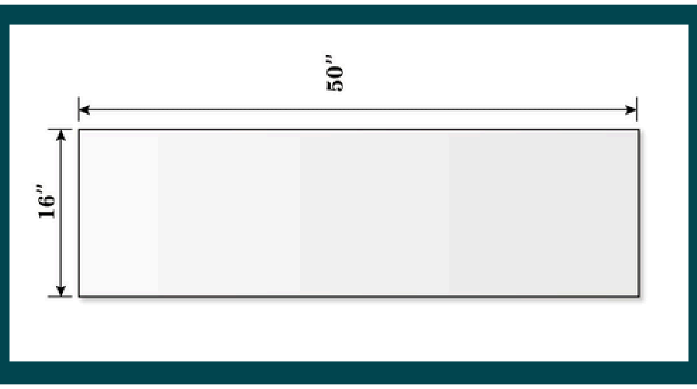
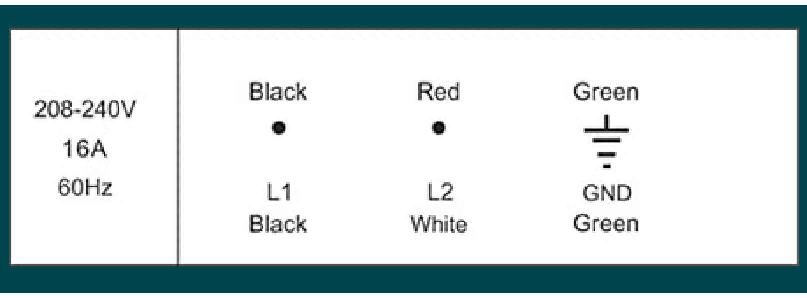
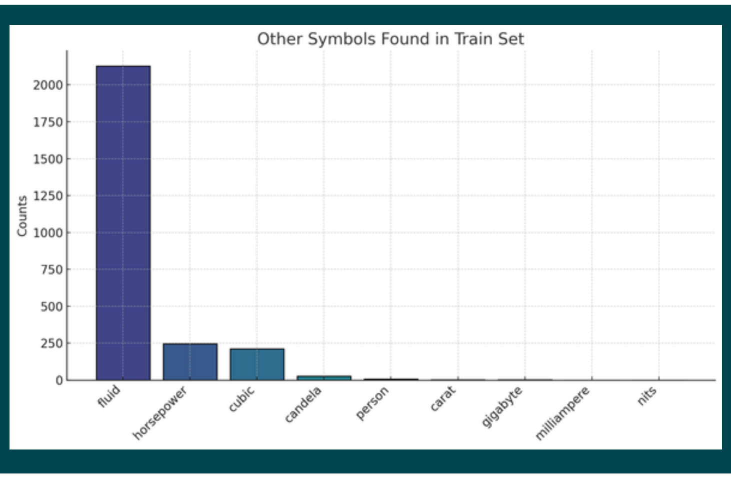
LLaMA-Factory
Easy and Efficient LLM Fine-Tuning

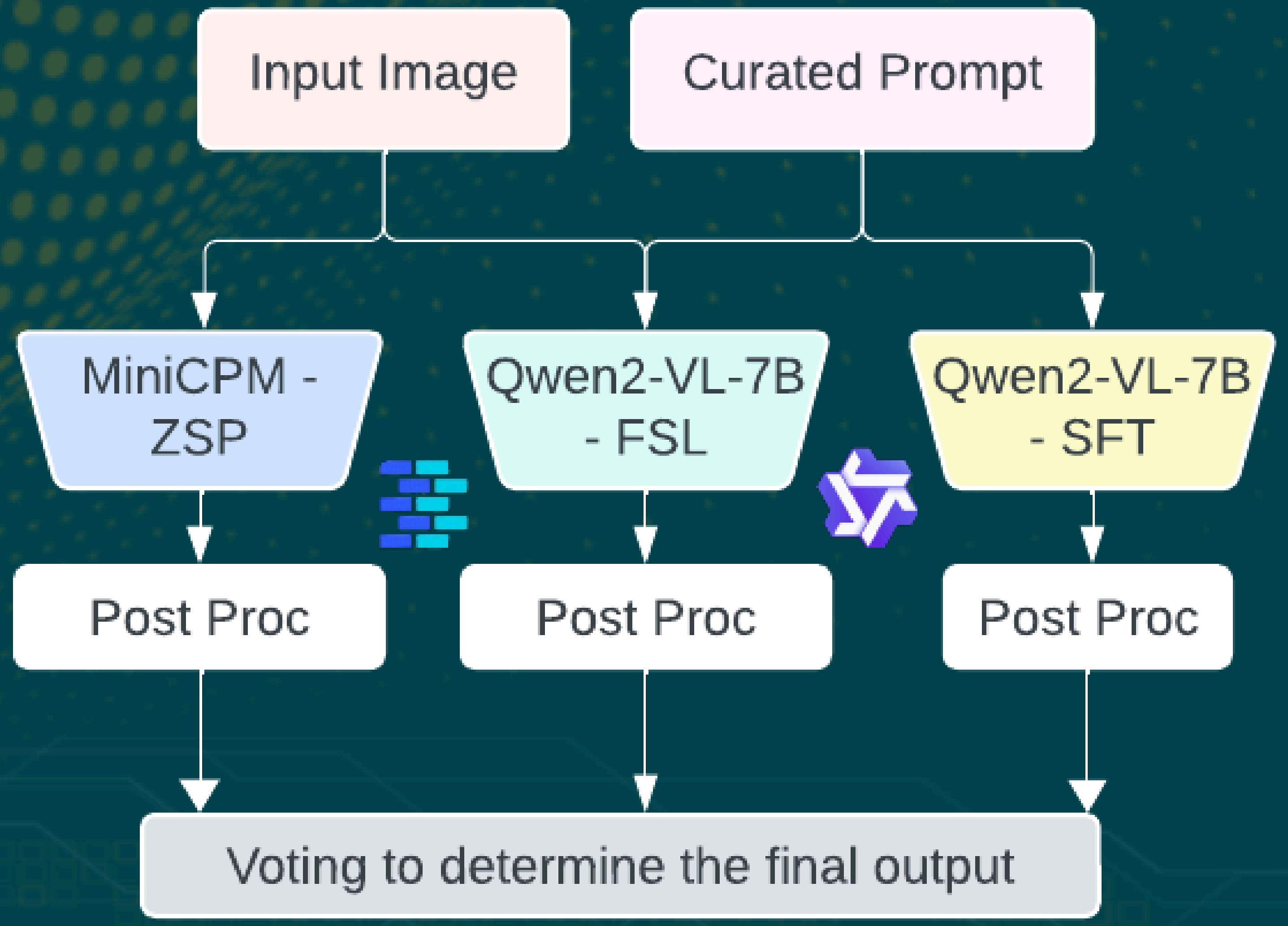


Method	Bits	7B	13B	30B	70B	110B	8x7B	8x22B
Full	AMP	120GB	240GB	600GB	1200GB	2000GB	900GB	2400GB
Full	16	60GB	120GB	300GB	600GB	900GB	400GB	1200GB
Freeze	16	20GB	40GB	80GB	200GB	360GB	160GB	400GB
LoRA/GaLore/BAdam	16	16GB	32GB	64GB	160GB	240GB	120GB	320GB
QLoRA	8	10GB	20GB	40GB	80GB	140GB	60GB	160GB
QLoRA	4	6GB	12GB	24GB	48GB	72GB	30GB	96GB
QLoRA	2	4GB	8GB	16GB	24GB	48GB	18GB	48GB

GPU VRAM vs Method

Post-Processing Model Output

Edge Cases in the Data	Example of the Edge Case	Solution
Fractions and Mixed Fractions in Images		• Regex expressions to extract fraction strings and convert into decimals
Use of symbols like single quote (') for feet and double quotes (") for inches		• Converted symbols to feet and inches as training set and mapping did not consist of such symbols.
Ranges given in the samples with '-' or from value 'a' to value 'b'	 208-240V 16A 60Hz Black Red Green L1 Black L2 White GND Green	• As per guidelines shared with participants, we took higher values in these cases.
Symbols other than those in the appendix were found in images		• Instruction-tuned and utilised few-shot learning to allow model to say this symbol is not in the list. • Rule-based algorithm to remove symbols.



Input Image

Curated Prompt

MiniCPM - ZSP

Qwen2-VL-7B - FSL

Qwen2-VL-7B - SFT

Post Proc

Post Proc

Post Proc

Voting to determine the final output

Final Pipeline - Ensemble

Scalability:

- Plug-and-play for any VLM
- Training ~ 3.5 hours running 4 A6000s
- Few-Shot ~ 3 hours on 20 GB VRAM

Future Improvements:

- Better filtering of data before SFT
- More GPU compute ~ better training

Strategy	Model	F1 Score
ZSP	MiniCPM-2.6	**66.2**
	InternVL2-8B	65.9
ZSP + PostProc	MiniCPM-2.6	**69.3**
	InternVL2-8B	68.2
SFT	Qwen2-7B-SFT	64.8
FSL	Qwen2-7B	**70.9**
Ensemble Methods	ZSP-1 + ZSP-2	68.5
	SFT + ZSP-1	70.7
	SFT + FSL	71.4
	SFT + FSL + ZSP-1	**71.8**

Final Results Table