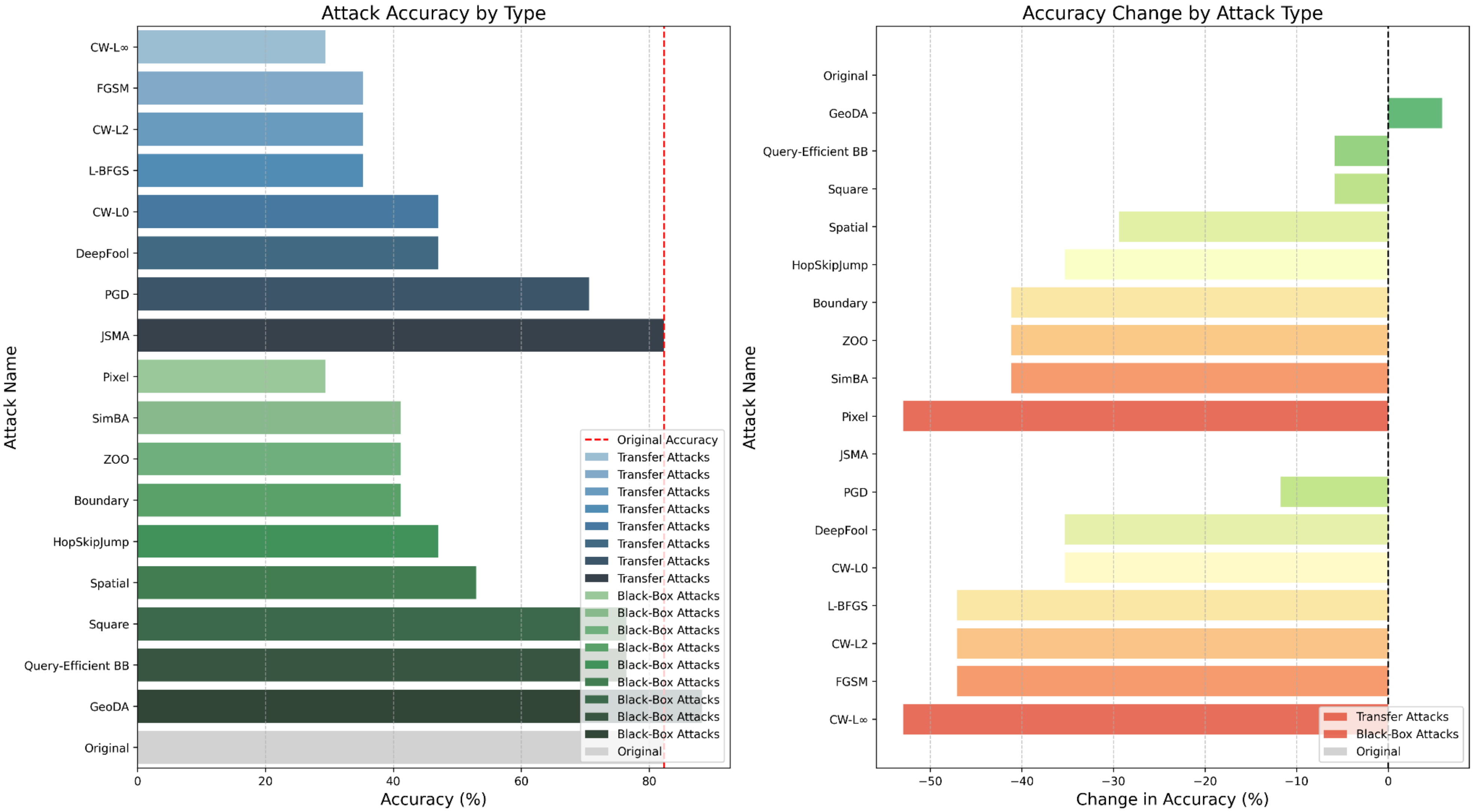
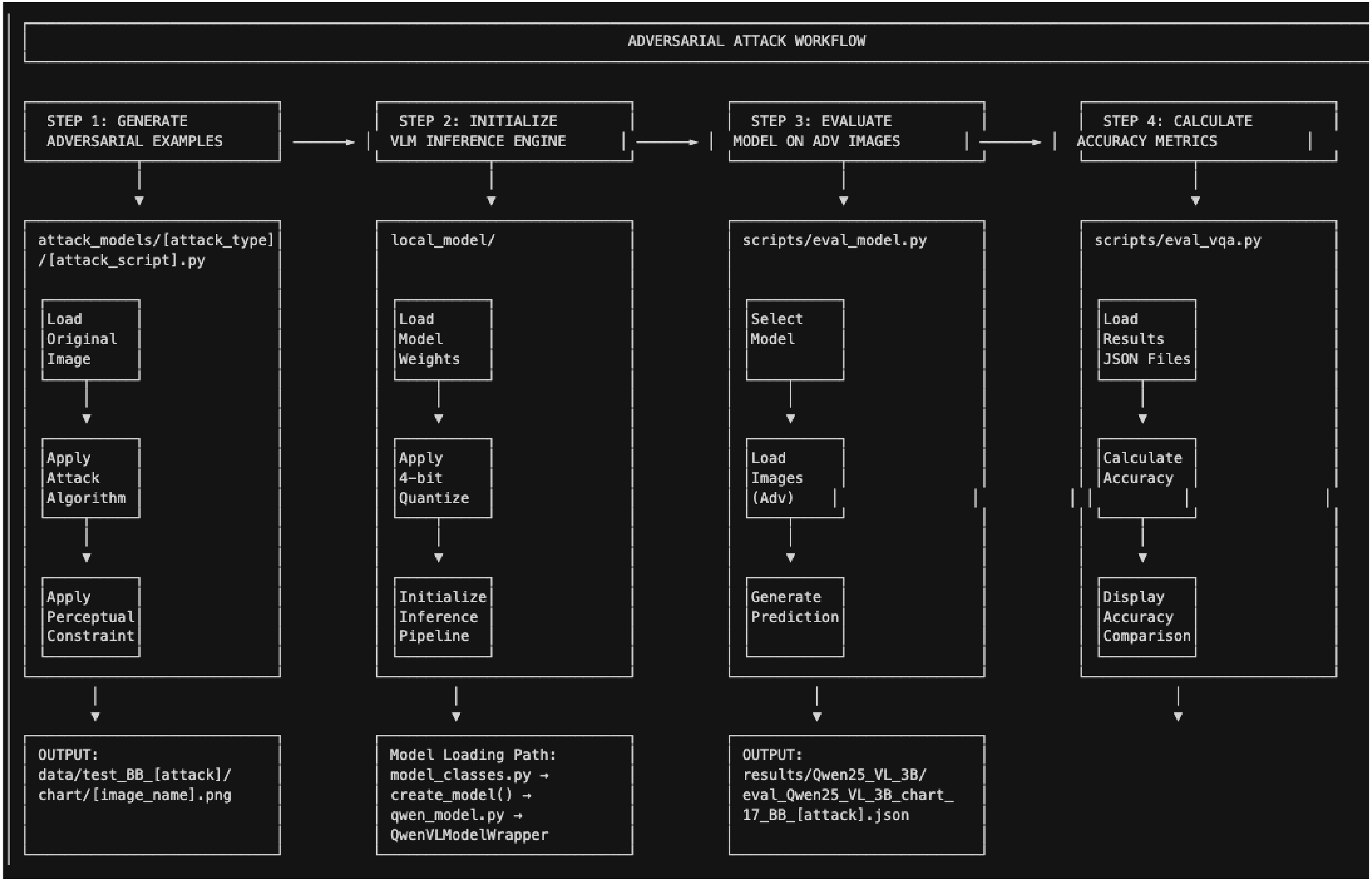

Attack Accuracy by Type
Attack Name
CW-L∞
FGSM
CW-L2
L-BFGS
CW-L0
DeepFool
PGD
JSMA
Pixel
SimBA
ZOO
Boundary
HopSkipJump
Spatial
Square
Query-Efficient BB
GeoDA
Original
0
20
40
60
80
Accuracy (%)
Original Accuracy
Transfer Attacks
Transfer Attacks
Transfer Attacks
Transfer Attacks
Transfer Attacks
Transfer Attacks
Transfer Attacks
Transfer Attacks
Black-Box Attacks
Black-Box Attacks
Black-Box Attacks
Black-Box Attacks
Black-Box Attacks
Black-Box Attacks
Black-Box Attacks
Black-Box Attacks
Black-Box Attacks
Original
Accuracy Change by Attack Type
Attack Name
Original
GeoDA
Query-Efficient BB
Square
Spatial
HopSkipJump
Boundary
ZOO
SimBA
Pixel
JSMA
PGD
DeepFool
CW-L0
L-BFGS
CW-L2
FGSM
CW-L∞
−50
−40
−30
−20
−10
0
Change in Accuracy (%)
Transfer Attacks
Black-Box Attacks
Original



ADVERSARIAL ATTACK WORKFLOW

STEP 1: GENERATE ADVERSARIAL EXAMPLES

attack_models/[attack_type]/[attack_script].py

Load Original Image

Apply Attack Algorithm

Apply Perceptual Constraint

OUTPUT:
data/test_BB_[attack]/chart/[image_name].png

STEP 2: INITIALIZE VLM INFERENCE ENGINE

local_model/

Load Model Weights

Apply 4-bit Quantize

Initialize Inference Pipeline

Model Loading Path:
model_classes.py →
create_model() →
qwen_model.py →
QwenVLModelWrapper

STEP 3: EVALUATE MODEL ON ADV IMAGES

scripts/eval_model.py

Select Model

Load Images (Adv)

Generate Prediction

OUTPUT:
results/Qwen25_VL_3B/eval_Qwen25_VL_3B_chart_17_BB_[attack].json

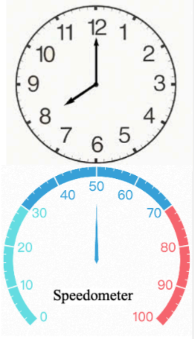
STEP 4: CALCULATE ACCURACY METRICS

scripts/eval_vqa.py

Load Results JSON Files

Calculate Accuracy

Display Accuracy Comparison

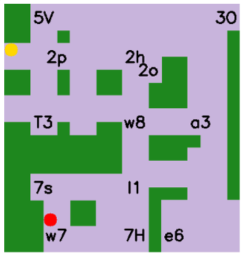


Dashboard

Question: Today, I set off for another city 100 kilometers away. The clock shows my departure time, and the speedometer displays my car's speed. When will I arrive at the destination?

Answer: 10:00 AM. **Rationale:** Your departure time is 8:00 AM, and your car's speedometer shows 50 km/h. The distance is 100 km. Therefore, you need to spend two hours on the road, and you will arrive at destination at 10:00 AM.

GPT-4V: You will arrive at 9:00. The clock shows 7:00, which is your departure time.. ✗



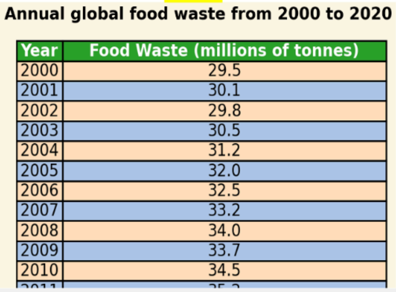
Road Map

Question: The starting point is marked in red, and the endpoint is in yellow. Please find a path from the starting to the endpoint.

Answer: Start from the red point, move upwards to 7s, and then move to the right to L1...

GPT-4o: 1.Move up to "w7"… ✗

Table



Annual global food waste from 2000 to 2020

Year	Food Waste (millions of tonnes)
2000	29.5
2001	30.1
2002	29.8
2003	30.5
2004	31.2
2005	32.0
2006	32.5
2007	33.2
2008	34.0
2009	33.7
2010	34.5

Question: How much food waste has increased from 2000 to 2009

Answer: 4.2 Rationale: 33.7 - 29.5

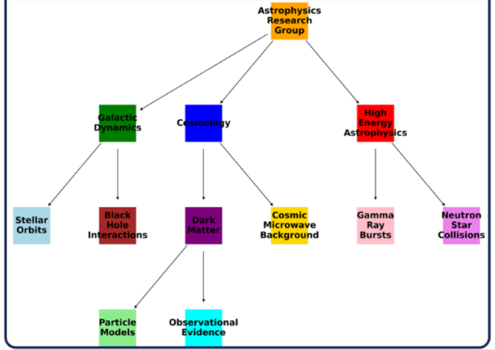
GPT-4V: 4.2 … increased from 29.5 million tonnes in 2000 to 33.7 million tonnes in 2009… ✓

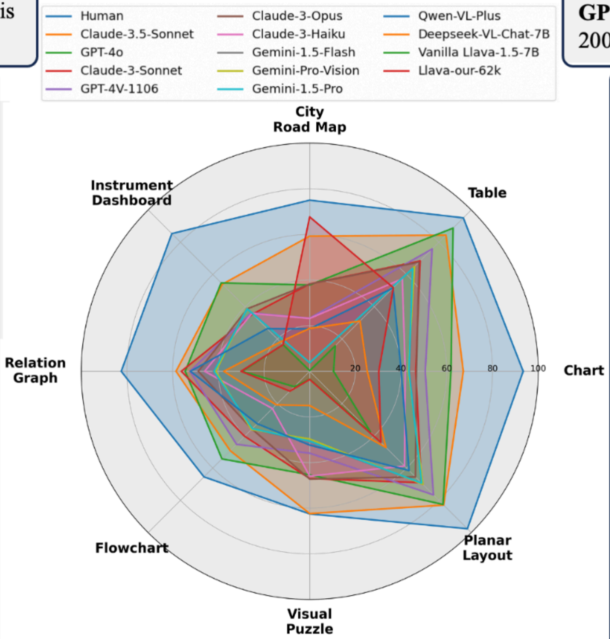
Relation Graph

Question: Do Stellar Orbits and Particle Models belong to the same sub-disciplines ? **Answer**: No.

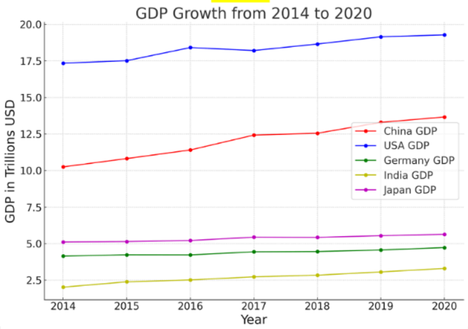
Rationale: Particle Models are a sub subfield of Dark Matter, while Stellar Orbits fall under Galactic Dynamics.

Llava-v1.6-34b: Yes, Stellar Orbits and Particle Models belong to the same sub-disciplines within the field of astrophysics … ✗





Chart



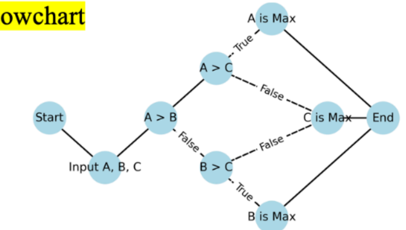
Question: How many times was the GDP of the USA in 2020 compared to China?

Answer: 1.3 times.

Rationale: In 2020, China's GDP was $14 trillion, while..

GPT-4V: 1.17 times, For China, the GDP is represented by the blue line, which intersects at about 15 trillion USD in 2020.… ✗

Flowchart

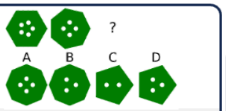


Question: According to this flowchart, I input the numbers 2, 3, and 4. What is the final result?

Answer: 4. **Rationale:** This algorithm flowchart is designed to find the maximum of three numbers

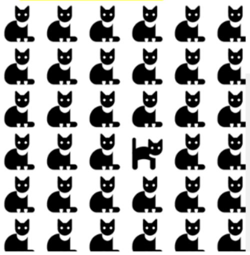
GPT-4o: C, which is 4, is the maximum value. Start with the comparison A > B (2 > 3), which is false …✓

Visual Puzzle



▲**Question:** Select an option to fill ?**Answer:** B. The sides are decreasing...

Llava-v1.6-34b: …without more clear pattern difficult to predict … ✗

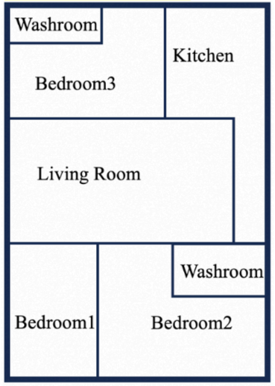


▲ **Question:** Observe the image. Are all the images identical?

Answer: No. **Rationale:** The cat in row 4 and column 4 is different …

GPT-4V: No, …unique cat is located in the fourth row and third column from the top left corner. ✗

2D Planar Layout



Question: Does the smallest bedroom have a washroom?

Answer: No.

Rationale: The smallest bedroom is BedRoom1, and it does not a washroom.

GPT-4V: Yes. Bedroom2.. ✗