# **Evaluating Vision-Language Models as Evaluators in Path Planning**

# Supplementary Material

# A. Scenarios

A complete of the 15 scenarios used in PATHEVAL are presented in Table 6. The descriptors that each scenario aims to optimize are listed in Table 7.

### **B.** Path Descriptors

For a path  $P = \{p_1, p_2, ..., p_n\}$  consisting of a sequence of *n* locations *p*'s, we define each of the descriptors as follows, where  $D(\cdot, \cdot)$  denotes the Euclidean distance between the two operands. The three clearance metrics and Path Length share the same measuring unit, i.e., one grid size; Smoothness and Maximum angle are measured by degree; and Number of Sharp Turns is an integer count.

1. *Minimum Clearance* measures the smallest distance between any point on the path and the nearest obstacle.

Min Clearance = 
$$\min_{p_j \in P} \left( \min_{O_i \in \mathcal{O}} D(p_j, O_i) \right)$$

2. *Maximum Clearance* measures the largest distance between any point on the path and the nearest obstacle.

Max Clearance = 
$$\max_{p_j \in P} \left( \min_{O_i \in \mathcal{O}} D(p_j, O_i) \right)$$

3. *Average Clearance* computes the average distance between all points on the path and the nearest obstacle.

Avg. Clearance = 
$$\frac{1}{n} \sum_{p_j \in P} \left( \min_{O_i \in \mathcal{O}} D(p_j, O_i) \right)$$

4. *Path Length* sums up the Euclidean distances between consecutive points on the path.

Path Length = 
$$\sum_{j=2}^{n} D(p_{j-1}, p_j)$$

5. *Smoothness* is defined as the sum of the angles between consecutive segments of the path, measuring how smoothly the path changes direction.

Smoothness = 
$$\sum_{j=3}^{n} \theta_j$$

where  $\theta_j$  is the angle between the vectors  $\overrightarrow{p_{j-2}p_{j-1}}$  and  $\overrightarrow{p_{j-1}p_j}$ .

6. *Number of Sharp Turns* counts the number of turns in the path where the angle between consecutive segments exceeds 90 degrees.

Sharp Turns = 
$$\sum_{j=3}^{n} \delta_j$$
, where  $\delta_j = \begin{cases} 1 & \text{if } \theta_j > 90^{\circ} \\ 0 & \text{otherwise} \end{cases}$ 

7. *Maximum Angle* denotes the largest angle between any two consecutive segments of the path.

Maximum angle = 
$$\max_{j=3}^{n} \theta_j$$

# **C. Experimental Setup**

# C.1. Prompts

Below, we show the different prompts used for our experiments, using Scenario #1 in Table 6 as an example.

### C.1.1 Prompt for PATHEVAL

### **Prompt for PATHEVAL**

The agent navigating this maze is a large truck, so sharp turns (90 degrees or larger) are harder to make. It should also stay on a straight line (unless it is making a turn) as it is driving through heavy traffic. Which path better achieves the task? Path 1 is on the left side and Path 2 is on the right side. Your answer should follow the format below: Answer: Path 1 or Path 2. Explanation: Why you chose the path (1 or 2).

### C.1.2 Prompt for PATHEVAL w/ Descriptors

### **Prompt for PATHEVAL w/ Descriptors**

The agent navigating this construction site is a long articulated bus, making it difficult to maneuver sharp turns (90 degrees or larger). Which path better achieves the task? Path 1 is on the left side and Path 2 is on the right side. The following path descriptor values are computed for each path:

**Minimum Clearance:** The minimum distance from the obstacles.

**Maximum Clearance:** The maximum distance from the obstacles.

**Smoothness:** The sum of absolute angles between path segments. Smoother paths have a lower smoothness value.

**Number of sharp turns:** Number of turns that are > 90 degrees.

Maximum turn angle: The sharpest turn angle in the path.

Path length: The sum of Euclidean distances between

ID	Scenario
1	The agent navigating this maze is a large truck, so sharp turns (90 degrees or larger) are harder to make. It should also stay on a straight line
	(unless it is making a turn) as it is driving through heavy traffic.
2	An autonomous firefighting robot is designed to navigate and operate within burning buildings to extinguish fires and rescue trapped individuals.
	It should explore as much of the area as possible, while maintaining a safe distance from the walls to avoid damage.
3	As the vehicle is traversing a warzone, it must stay concealed from enemy operatives, making use of covers like walls and avoiding open spaces
	as much as possible. It should also reach its target (point 2) as quickly as possible.
4	An autonomous drone delivering a package from point 1 to point 2 must take the shortest path possible due to limited fuel. It should also maintain
	a safe distance from surrounding buildings and make the path as straight as possible for stable flight.
5	A robot has to deliver an aid package from point 1 to point 2 as quickly as possible. As the vehicle is moving through an earthquake-affected area,
	it is crucial to keep a safe distance from the walls at every moment to prevent damage from collapsing structures.
6	A robot is moving through a museum where the walls contain fragile and expensive art pieces. Therefore, the robot should make sure to never get
	too close or touch any of the walls. It should also not take any abrupt turns to avoid startling the visitors.
7	The agent navigating this construction site is a long articulated bus, making it difficult to maneuver sharp turns (90 degrees or larger).
8	The agent navigating this trail is a wide agricultural combine harvester, making it difficult to see obstacles; hence it's hard to avoid them if they're
	too close.
9	The agent navigating this busy warehouse is a long forklift, making it difficult to make sharp and abrupt turns. It should also maintain a safe
	distance from the obstacles at all times.
10	The agent navigating this complex construction site is a crane with a long boom, which makes maneuvering sharp turns and around narrow
	passages very challenging.
11	An autonomous taxi is navigating through an urban environment. As it is navigating heavy traffic, it should make as few sharp turns as possible
	and keep a safe distance from its surroundings. It should also ensure passenger comfort and safety by making left/right turns as smooth as possible.
12	A Mars rover is exploring a Martian terrain from point 1 to point 2. The rover should conserve energy by taking the shortest path possible and
	avoiding unnecessary turns. Sharp turns (> 90 degrees) require higher levels of fuel and put a strain on the navigation system.
13	An autonomous vehicle is guiding a visually impaired individual through a shopping mall. It should drive in a straight path and not make any
	sudden or sharp turns to ensure the individual's safety and comfort. It should also maintain a safe distance from the surrounding walls.
14	An autonomous soil monitoring robot is tasked with navigating agricultural fields and collecting detailed soil health data. It should cover as much
	of the area as possible and get as close to the walls as possible to read the sensors that record the data needed.
15	An autonomous inspection robot is tasked with navigating a nuclear power plant to inspect for radiation leaks and structural integrity. The robot
	has to inspect as many sections of the power plant as possible in one mission. It should get as close as possible to the walls to be able to detect
	minor leaks or cracks. In order to avoid accidents, it should take the straightest path possible and not make any sudden or sharp turns.

Table 6	Scenario Descriptions with Corresponding IDs	

Scenario	Avg. clearance	Min. clearance	Max. clearance	Path length	#of Sharp turns	Max angle	Smoothness
1	-	-	-	-	$\downarrow$	$\downarrow$	$\downarrow$
2	1	$\uparrow$	-	$\uparrow$	-	-	-
3	$\downarrow$	-	$\downarrow$	$\downarrow$	-	-	-
4	$\uparrow$	$\uparrow$	-	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
5	-	$\uparrow$	-	$\downarrow$	-	-	-
6	-	$\uparrow$	-	-	$\downarrow$	-	$\downarrow$
7	-	-	-	-	$\downarrow$	$\downarrow$	$\downarrow$
8	$\uparrow$	$\uparrow$	-	-	-	-	-
9	-	$\uparrow$	-	-	$\downarrow$	$\downarrow$	$\downarrow$
10	-	-	-	-	$\downarrow$	$\downarrow$	$\downarrow$
11	$\uparrow$	$\uparrow$	-	-	$\downarrow$	$\downarrow$	$\downarrow$
12	-	-	-	$\downarrow$	$\downarrow$	-	$\downarrow$
13	$\uparrow$	$\uparrow$	-	-	$\downarrow$	$\downarrow$	$\downarrow$
14	$\downarrow$	-	$\downarrow$	$\uparrow$	-	-	-
15	$\downarrow$	-	$\downarrow$	$\uparrow$	$\downarrow$	$\downarrow$	$\downarrow$

Table 7. Scenarios and the descriptors over which they optimize.  $\downarrow$  indicates that values need to be minimized, while  $\uparrow$  means that values have to be maximized.

points in the path.

Here are path descriptor values for Path 1: **Minimum clearance:** 0.7044694115091165, **Max-** imum clearance: 6.142489571740198, Average clearance: 3.014227976325727, Path length: 137.5945426777758, Smoothness: 98.5529683186464, Sharp turns: 0, Maximum angle: 59.600200981198626.

Here are path descriptor values for Path 2: **Minimum clearance:** 1.342789990448996, **Maximum clearance:** 5.343046965766502, **Average clearance:** 2.697627901518315, **Path length:** 152.13523628046815, **Smoothness:** 480.5614409019347, **Sharp turns:** 1, **Maximum angle:** 108.15343849689171.

Your answer should follow the format below: **Answer**: Path 1 or Path 2. **Explanation**: Why you chose the path (1 or 2).

### C.1.3 Prompt for Attribute Abstraction

In the following, we present the prompt we used to understand a VLM's attribute abstraction capability (Appendix D.1). The experiment queries each VLM to decide which descriptors are important for the given scenario.

### **Prompt for Attribute Abstraction**

The agent navigating this maze is a large truck, so sharp turns (90 degrees or larger) are harder to make. It should also stay on a straight line (unless it is making a turn) as it is driving through heavy traffic.

The following descriptors are available:

- 1. **Minimum Clearance**: The minimum distance from the obstacles.
- 2. Maximum Clearance: The maximum distance from the obstacles.
- 3. Average Clearance: The average distance from the obstacles.
- 4. **Smoothness**: The sum of absolute angles between path segments. Smoother paths have a lower smoothness value.
- 5. Number of Sharp Turns: The number of turns that are >90 degrees.
- 6. Maximum Turn Angle: The sharpest turn angle in the path.
- 7. **Path Length**: The sum of Euclidean distances between points in the path.

Which ones are the most important for the specified scenario?

### Your answer should follow this format:

**Answer**: list of required descriptors separated by a semicolon (;).

Explanation: Why these descriptors are important.

### C.1.4 Prompt for Fine-grained Visual Perception

# Prompt for Fine-grained Visual Perception

Smoothness is defined as a measure of how gradual the agent's path is, minimizing sharp or abrupt changes in direction. It is calculated as the sum of angles between consecutive points (segments).

The task is to determine which path results in a numerically smaller value for smoothness. A smaller smoothness value means that the path has fewer abrupt turns and is smoother overall.

Path 1 is on the left side and Path 2 is on the right side. **Your answer should follow this format:** 

- Answer: Path 1 or Path 2.
- **Explanation**: Briefly explain why you chose the path (e.g., "Path 1 has a smaller value for the given metric").

# **C.2. Fine-grained Vision Dataset Construction**

Intuitively, when the difference between two paths increases, it should become easier for a VLM to pick the one with a lower value. To create such a controlled dataset, for each descriptor m we consider three threshold values  $\epsilon_1, \epsilon_2, \epsilon_3$ , where  $\epsilon_1 < \epsilon_2 < \epsilon_3$ , and then sample 50 path pairs  $\{(P_i, P_j)\}$  for each threshold  $\epsilon_i$  such that  $|m(P_i) - m(P_j)| > \epsilon_i$ . We list the threshold values in Table 8 and report the resulting average descriptor value for each subset in Table 9.

Descriptor	$\epsilon_1$	$\epsilon_2$	$\epsilon_3$
Min. Clearance	1.0	2.0	3.0
Max. Clearance	1.0	2.0	3.0
Avg Clearance	1.0	2.5	5.0
Path Length	50	75	100
Smoothness	100	200	300
# of Sharp Turns	1	2	3
Max. Angle	30	60	90

Table 8. Fine-grained visual perception thresholds used in our analysis for understanding a VLM's low-level perception.

Descriptor	Avg. 1	Avg. 2	Avg. 3
Min. Clearance	1.7	2.7	3.7
Max. Clearance	2.4	3.3	3.8
Avg Clearance	2.4	4.2	5.8
Path Length	72.4	91.2	113.4
Smoothness	339.4	372.5	447.6
# of Sharp Turns	2.6	3.5	4.3
Max. Angle	57.6	74.7	98.7

Table 9. Average differences in descriptor values of path pairs in our synthesized dataset for understanding a VLM's low-level perception.

For comparison, we also show the average difference between path pairs on PATHEVAL for each descriptor in Table 10.

Descriptor	Avg Difference
Minimum clearance	0.593
Maximum clearance	0.124
Average clearance	1.31
Path length	19.72
Smoothness	140.93
# of Sharp turns	0.582
Maximum angle	32.31

Table 10. Average differences in PATHEVAL.

### C.3. Experimental setup for VLM fine-Tuning

In Table 11, we present the hyper-parameters used in finetuning LLaVA-NeXT-7b. Our experiments were performed using A100.80gb GPUs.

Language Decoder	Vicuna-7b [13]
Vision Encoder	clip-vit-large-patch14-336
Number of Epochs	50
Batch size	16
Learning Rate	2e-5

Table 11. Hyper-parameters used for fine-tuning.

# **D.** Additional Analysis

### **D.1. Can VLMs identify relevant descriptors?**

We assess whether VLMs are able to identify the relevant descriptors for each scenario. Accordingly, we prompt the models with each scenario and the list of available descriptors and ask the model to identify which ones are important for the scenario (using the prompt template in Appendix C.1.3). We run this process 5 times for each of the 15 scenarios. Table 12 shows the average number of times *at least one of the* relevant descriptors are chosen by each model across all instances. (For the path pairs in PATHE-VAL, knowing one of the critical descriptors and being able to identify the path with a smaller or larger value as desired are sufficient for succeeding in the task.)

Model	Avg. Success
GPT-40-mini	1.00
GP1-40	1.00
Qwen2-VL-7b	1.00
LLaVA-OneVision-7b	1.00
LLaMa-3.2-11b	0.92
LLaVa-NeXT-13b	0.93

Table 12. Average number of times *at least one of the* relevant descriptors are chosen by each model across all scenarios.



Figure 3. GPT-40 performance per scenario (2D)

# D.2. Some decision-making scenarios are easier than others.

In other to get a better understanding of how the models utilize the provided information, we look at the performance gain in each of the 15 scenarios for the best-performing model (GPT-40). This performance is broken down in Figure 3. We notice that GPT-40 may overuse its commonsense. Specifically, we observed that the model particularly struggled with Scenario #2 (Table 6). This is because, contrary to most real-case scenarios, this case requires maximizing the path length, to be able to cover as much of the area as possible. However, as finding the shortest path is generally a more common command, GPT-40 has likely been pre-trained to prioritize this pattern and thus often opts for paths minimizing the length on PATHEVAL. On the other hand, GPT-40 was also shown to struggle in using its commonsense knowledge properly. Specifically, its performance on Scenario #8 does not improve even when the descriptors are listed. This is because, in this scenario, the specification does not explicitly list the aspects that the model has to consider. This indicates that the models struggle to use their commonsense reasoning to identify aspects that are not explicitly stated in the specifications.

### D.3. Representations produced by the vision encoders are ambiguous.

One challenge encountered with encoders such as CLIP occurs is that similar images are represented almost identically [63, 64]. To investigate whether this is the issue faced in our case, we examine whether the model can easily memorize a small set of samples when fine-tuned for 100 epochs. If the representations were fully unambiguous, this memorization task should be straightforward. The results, displayed in Table 13, indicate that the model struggles to differentiate between several image and scenario pairs, as reflected by its low performance in the standard setting. Notably, performance significantly improves when descriptors or even random tokens are added to the language descriptions, making each image-language pair unique.

# of Instances	Image Only	Image + Desc.	Image + Rand. Tokens
50	0.72	1.0	1.0
100	0.68	1.0	1.0
200	0.77	1.0	1.0

Table 13. Fine-tuned LLaVA-NeXT-7b (w/ frozen vision encoder)'s accuracy in memorizing small training samples (2D) when each path is represented by only its image, image with descriptor values, or image with random text tokens (used to validate the VLM's small-data memorization capability).

# **E. Failure Examples**

We present some examples of failures produced by different models in Figures 4 and 5.

# F. Human Performance on the Task:

We sampled a set of 60 examples (2 from each scenario in both 2D and 3D) from PathEval and tasked four annotators with selecting the better path. Two of the annotators are authors of the paper (Ann. 3&4) and the other two are externals unfamiliar with the work (Ann. 1&2). We notice that in all cases the human annotators significantly outperform GPT-40 on the sampled subset, however, we notice that in the 3D case, the task can be difficult even for humans.

Ann. 1	Ann. 2	Ann. 3	Ann. 4	GPT-40
0.90/0.70	0.73/0.60	0.90/0.76	0.80/0.73	0.53/0.47

### Failure Examples on PATHEVAL



#### Prompt

The agent navigating this maze is a large truck, so sharp turns (90 degrees or larger) are harder to make. It should also stay on a straight line (unless it is making a turn) as it is driving through heavy traffic. Which path better achieves the task? Path 1 is on the left side and Path 2 is on the right side.

Your answer should follow the format below: Answer: Path 1 or Path 2.

Explanation: Why you chose the path (1 or 2).

#### GPT-40

#### Answer: Path 2

Explanation: Path 2 has fewer sharp turns compared to Path 1, making it more suitable for a large truck that has difficulty making sharp turns. Additionally, Path 2 allows the truck to stay on a straight line for longer distances, which is beneficial for navigating through heavy traffic.

#### LLaMA-11b Answer: Path 2

#### **Explanation**:

Step 1: Identify the key constraints of the task.

The agent is a large truck, which means it has difficulty making sharp turns (90 degrees or larger). Additionally, the agent should stay on a straight line as much as possible due to heavy traffic.

Step 2: Analyze Path 1.

Path 1 involves several sharp turns, including a 90-degree turn. This would be challenging for the large truck to navigate.

nStep 3: Analyze Path 2.

Path 2 involves fewer sharp turns and more gradual turns, which would be easier for the large truck to navigate.

Step 4: Determine the best path based on the constraints.

Given the constraints, Path 2 is the better option as it involves fewer sharp turns and more gradual turns, making it easier for the large truck to navigate.

#### LLaVA-OneVision-7b

#### Answer: Path 2

Explanation: Path 2 is a straight line from the starting point to the end point, which is the most efficient route for a large truck with difficulty making sharp turns. It avoids any sharp turns and stays on a straight line, which is ideal for driving through heavy traffic.

Figure 4. Examples of model failure on PATHEVAL.

#### Failure Examples on PATHEVAL w/ Descriptors



#### Prompt

The agent navigating this construction site is a long articulated bus, making it difficult to maneuver sharp turns (90 degrees or larger). Which path better achieves the task? Path 1 is on the left side and Path 2 is on the right side. The following path descriptor values are computed for each path:

Minimum Clearance: The minimum distance from the obstacles.

Maximum Clearance: The maximum distance from the obstacles.

Smoothness: The sum of absolute angles between path segments. Smoother paths have a lower smoothness value.

*Number of sharp turns: Number of turns that are > 90 degrees.* 

*Maximum turn angle:* The sharpest turn angle in the path.

Path length: The sum of Euclidean distances between points in the path.

Here are path descriptor values for Path 1:

Minimum clearance: 0.7044694115091165, Maximum clearance: 6.142489571740198, Average clearance: 3.014227976325727, Path length: 137.5945426777758, Smoothness: 98.5529683186464, Sharp turns: 0, Maximum angle: 59.600200981198626.

Here are path descriptor values for Path 2:

Minimum clearance: 1.342789990448996, Maximum clearance: 5.343046965766502, Average clearance: 2.697627901518315, Path length: 152.13523628046815, Smoothness: 480.5614409019347, Sharp turns: 1, Maximum angle: 108.15343849689171.

Your answer should follow the format below: Answer: Path 1 or Path 2.

Explanation: Why you chose the path (1 or 2).

### LLaVA-NeXT-7b

#### Answer: Path 2

**Explanation**: Path 2 is chosen because it has a **lower smoothness value**, which indicates that it has **fewer sharp turns** and a more gradual path. This is advantageous for a long articulated bus, as it would be easier to maneuver and avoid obstacles without getting stuck or damaging the bus. **The higher clearance values for Path 2** also suggest that it is less likely to encounter issues with obstacles.

### LLaVA-NeXT-13b

### Answer: Path 2

**Explanation:** Path 2 is chosen because it has a **lower smoothness value**, which indicates that it is a smoother path. This is important for a long articulated bus, as it will require less maneuvering and turning to navigate the construction site. The path length is also slightly shorter, which could be advantageous in terms of fuel efficiency and ease of navigation. Additionally, Path 2 has a higher minimum clearance, which is important for the bus to avoid hitting any obstacles. The maximum clearance is also higher, providing more margin for error. The only downside is that it has one more sharp turn compared to Path 1, but given the other advantages, it is still a more suitable path for the articulated bus.

#### InternVL2-8b

#### Answer: Path 2

**Explanation**: Path 1 has a higher number of sharp turns (0 compared to 1) and a higher smoothness value (98.5529683186464 compared to 480.5614409019347), which suggests that it is a smoother path overall. However, Path 1 also has lower clearance values compared to Path 2. Moreover, Path 1 has a smaller path length (137.5945426777758 compared to 152.13523628046815), which indicates a shorter path that still manages to navigate through the obstacles effectively. On the contrary, Path 2, although it doesn't have as many sharp turns, has better parameters like Maximum Clearance (5.343046965766502), Average clearance (2.697627901518315) which are indicators for the agent like a long articulated bus to maneuver safely. (Wrong descriptors are chosen)

Figure 5. Examples of model failure on PATHEVAL when prompted with w/ descriptor values.