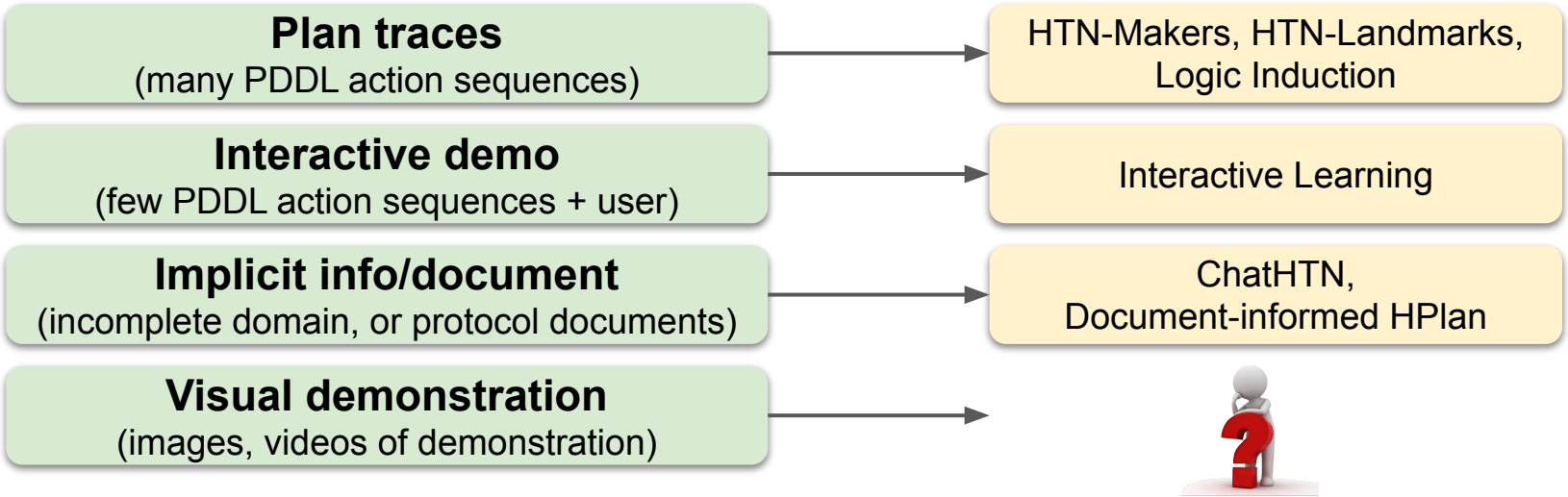
# Learning HTNs from Visual Demonstration with VLMs

Ngoc La, Karthik Mahadevan, Pulkit Verma, Julie A. Shah

# Introduction

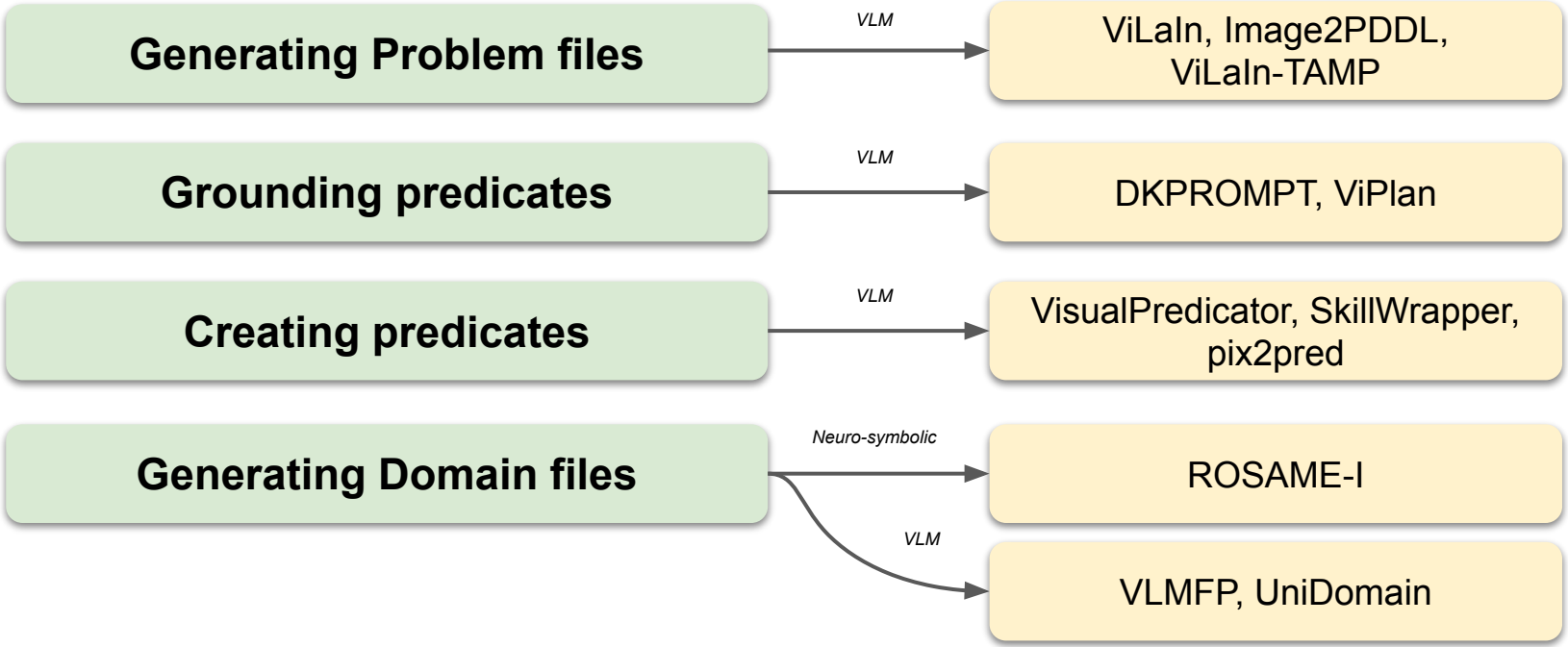
# Learning HTNs

- Existing HTN learning methods reduce manual knowledge engineering, but to my knowledge, learning explicit HTN structure directly from **raw visual demonstrations** remains underexplored.



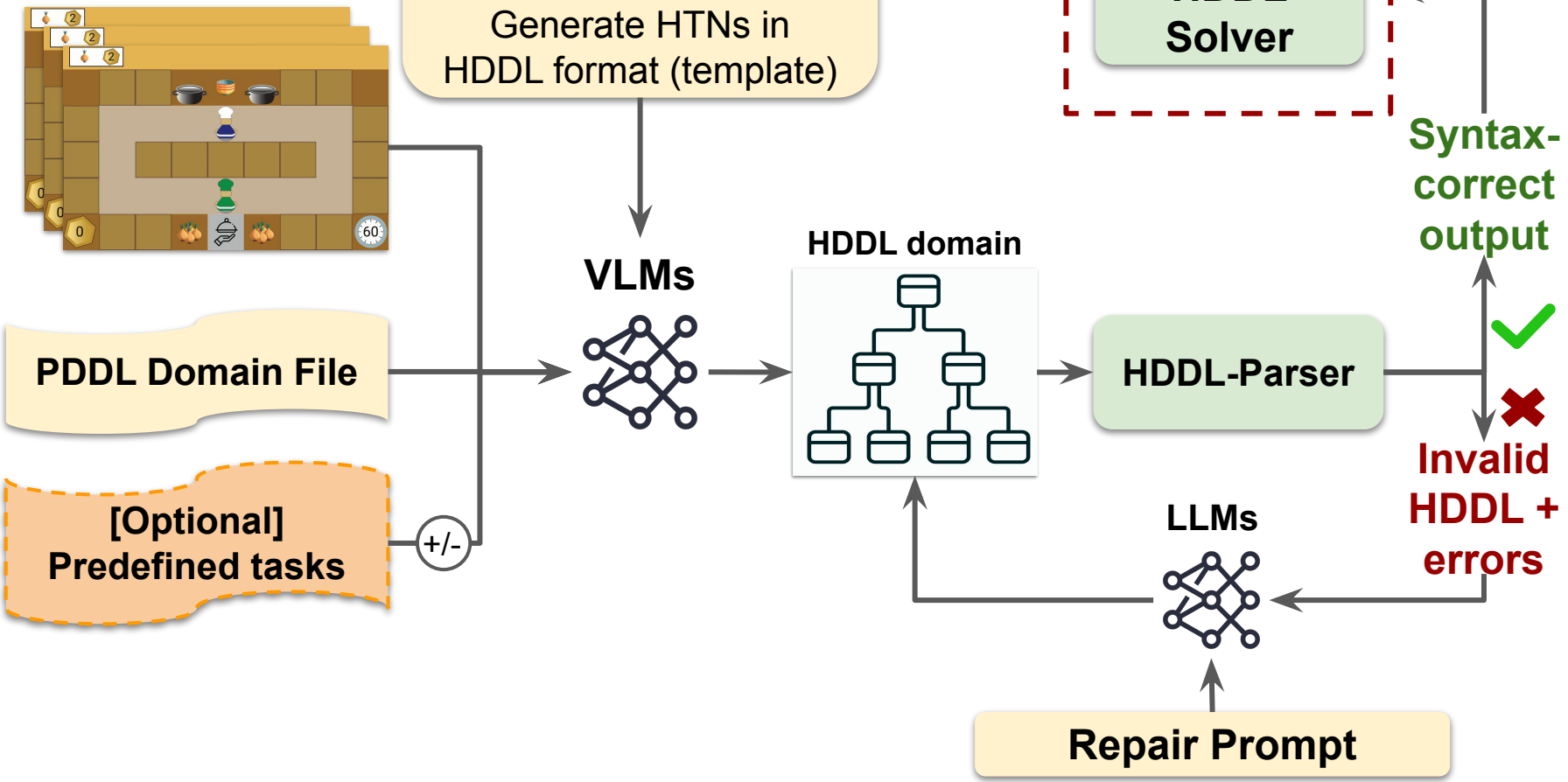
# Learning Symbolic Models from Visual Input

- Research has investigated construction of symbolic planning representations (PDDL) from visual input



# Vision2HTN Framework

- **VLMs** have shown strong capabilities in **interpreting images and videos**
- Investigate their potential for learning HTNs from visual demonstrations, a natural modality through which humans provide task knowledge.


Prompt:
Given visual data,
Given PDDL domain,
Generate HTNs in
HDDL format (template)
Evaluation
HDDL
Solver
Syntax-
correct
output
PDDL Domain File
[Optional]
Predefined tasks
+/-
VLMs
HDDL domain
HDDL-Parser
Invalid
HDDL +
errors
LLMs
Repair Prompt

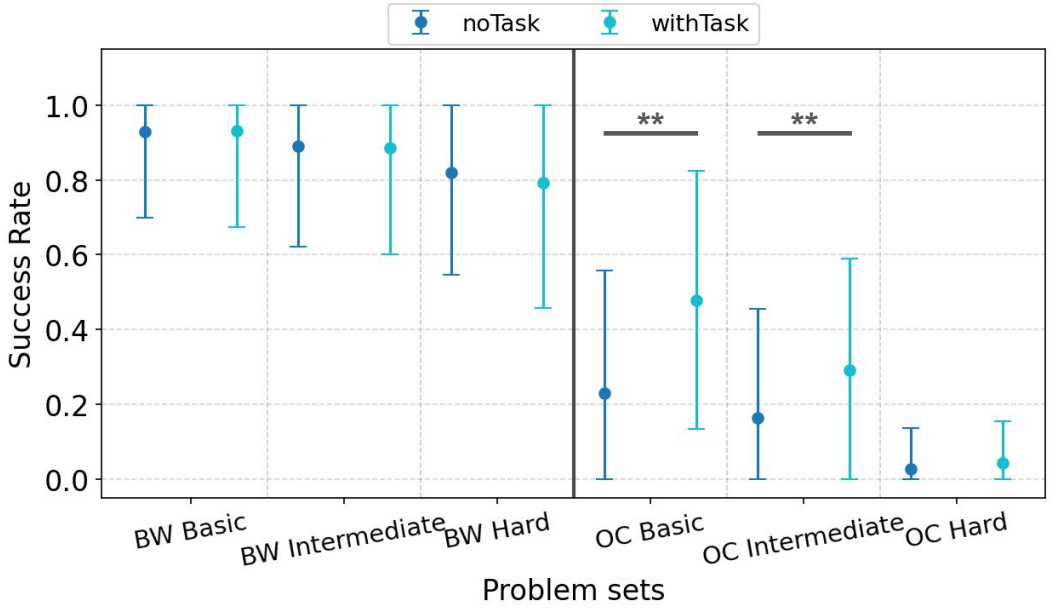
# Experiments

- Three factors:
  - Presence of predefined tasks
  - Presence of visual data:
    - No visual data (null)
    - Basic visual data
    - Intermediate visual data
    - Hard visual data
  - Choice of VLM:
    - OpenAI's GPT 5.4
    - Claude Sonnet 4.6
    - Gemini 3 Pro

# Results & Discussion

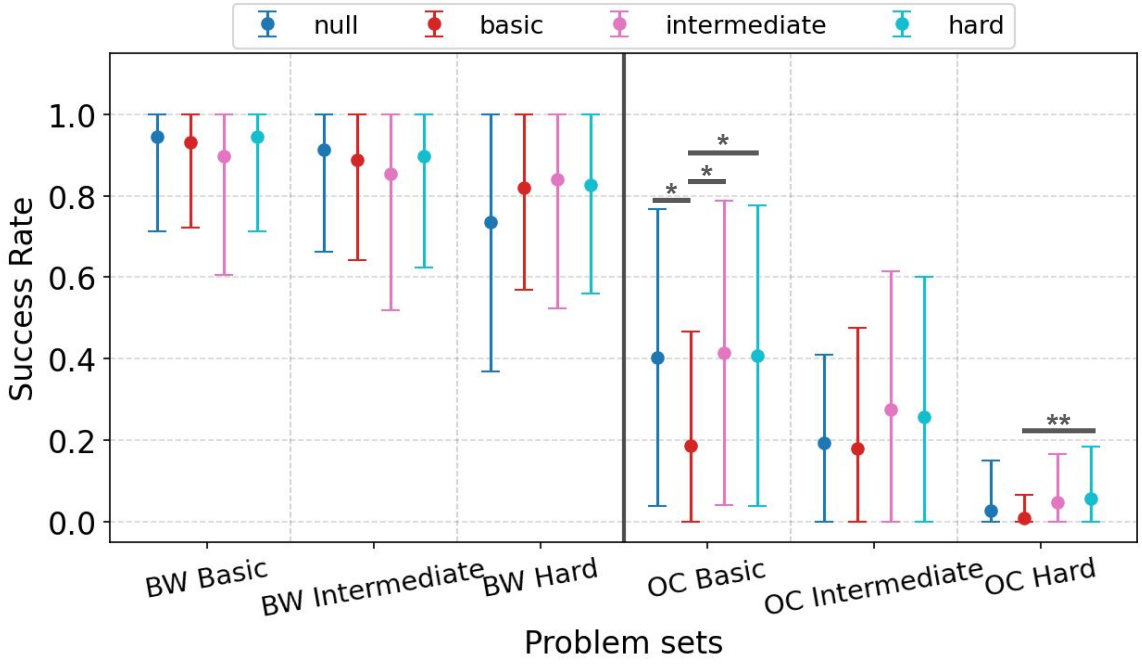
# Evaluate Vision2HTN: presence of task information

- Blocksworld: no clear difference as it is a popular domain that VLMs may be pre-trained with.
- Overcooked: providing information of task help generated better hierarchical strategies.



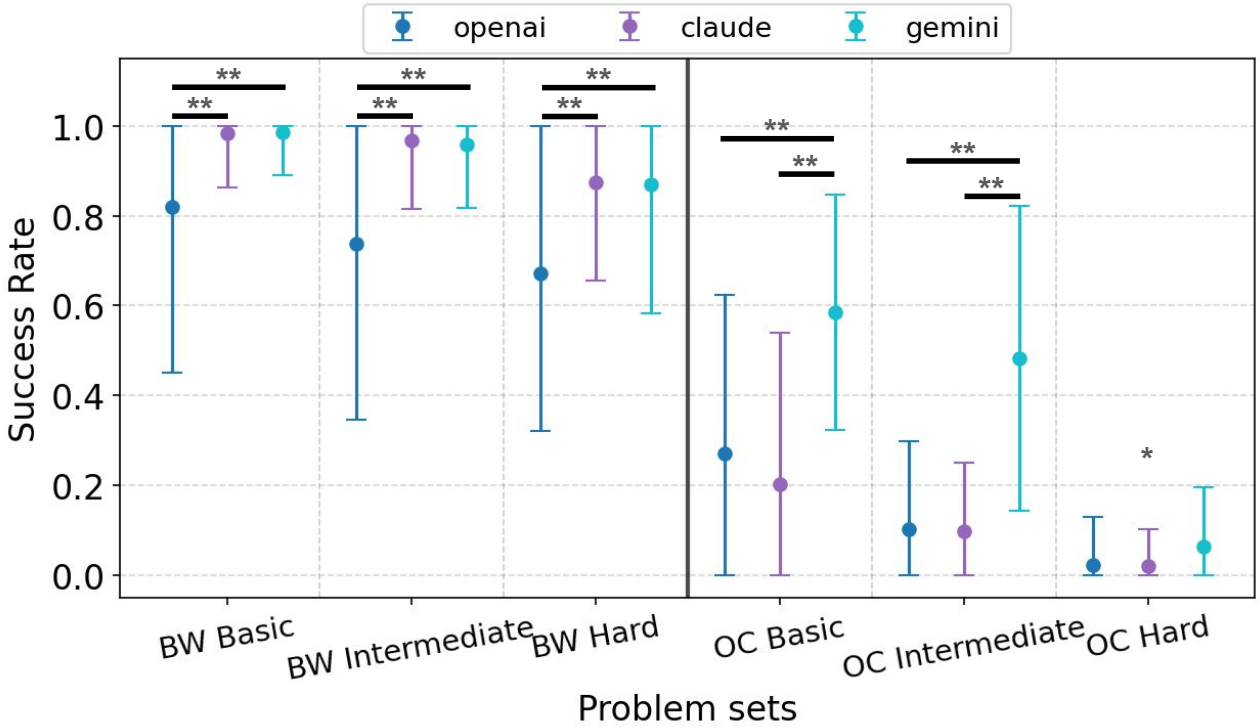
# Evaluate Vision2HTN: visual data mode

- Blocksworld: adding visual data add constraints to generating domain, thus perform less effective.
- Overcooked: visual data is more useful in more complicated problems



# Evaluate Vision2HTN: choice of VLMs

- Blocksworld: Claude perform the best
- Overcooked: Gemini performs the best



Thank you!