



Self-Skill-Exploration:



Tools



Sub-Skills /
Tool functions



Easy
Question



Hard
Question



Skills



Sub-skill Library

Input Layer:
<option_discriminative_analysis>
<procedural_knowledge_retrieval>
<question_structure_parsing>

Perception Layer:
<clip_video_segmentation>
<clip_text_visual_scoring>
<object_detection>
<temporal_reference_generation>

Decision Layer:
<vlm_step_identification>
<vlm_evidence_augmented_answer>
<vlm_text_reasoning>
<fuzzy_text_matching>
<set_elimination>

Tool Library



<GroundingDINO>
<Object detection based on text in options>
<Output: Confidence score, BBox>



<Clip>
<Computing the similarity between video
content and textual descriptions.>



<Web Search WikiHow>
<Constructing a knowledge base by retrieving
relevant information from WikiHow.>



<Temporal Segmentation>
<Perform temporal segmentation based on the
sequence of steps and time stamps.>



Qwen3.5-Plus as Brain



Question types
+ Options

<question_type 1>
...
<question_type n>

For each **question**

Exploration Round 1: Learn to solve easy questions first

Skill 1

if skill1(.) = skill2(.),
tag <easy question>
else
tag <hard question>

Skill 2

Exploration Round 2: Solve hard questions, learn from easy questions

<easy question>
skill1(), skill2() as
reference

<hard question>

Skill 1

Skill 2

<question_type x>

<question 1> <skill1> <skill2>
...
<question n> <skill1> <skill2>

For each **question_type**

Top-2 skill

Exploration Round 3: Summarize

<skill1>

<skill2>

+

<question 1> <skill1> <skill2>
...
<question n> <skill1> <skill2>

Reasoning
Summarize

Top-2 skills for <question_type x>

Skill 1

Skill 2

Execute