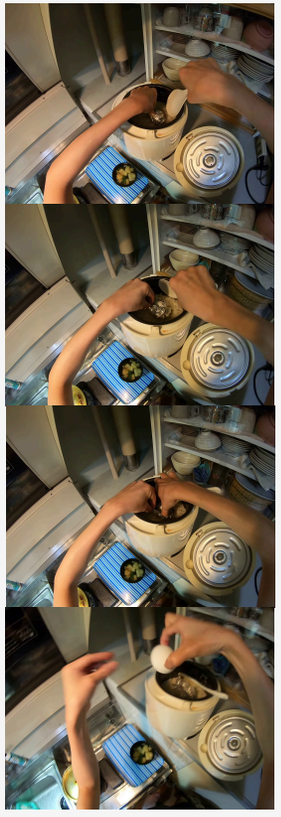
Stage 1

Stage 2

Stage 3

VIDEO FRAMES



MLLM

[CAPTION]
#C C scoop out the paper from the rice cooker.

#C C move paper towards the sink.

#C C dispose of paper in the sink.

Planning Prompt

You are a visual question answering expert. You must choose .....

[OBSERVATION]

[QUESTION]
The task goal is Scoop out the paper from the rice cooker and put it in the sink. What action is C about to take?

[CAPTION]
#C C scoop out the paper from the rice cooker .....

[OPTION]
OPTION 0: put serving spoon in rice cooker.
OPTION 1: place egg.
OPTION 2: pick up egg from rice cooker.
OPTION 3: scoop paper in rice cooker.

MLLM

Decision Prompt

Given two answers, you must choose the option that is more reasonable .....

[OBSERVATION]

ANSWER 1

[DESCRIPTION]:
Based on the observation, C is holding an egg over the counter .....

[ANSWER]: 1

ANSWER 2

[DESCRIPTION]:
In the current observation, C's hands are interacting with a kitchen sink area .....

[ANSWER]: 0

MLLM

Final Answer

[REASON]:
ANSWER 1 is more reasonable because it accurately reflects the observation and details .....

[ANSWER]: 1