Given Textual Input

What's the step-by-step procedure for How to cook kimchi fried rice (<task>) ?

LLM

Vanilla Text Plan

Step 1
text: "First, you need to gather all your ingredients"
context: "You will need cooked rice (preferably cold or day-old), vegetables of your choice (e.g., carrots, ... "

Step 2
text: "Next, heat a wok over medium-high heat."
context: "When the pan is hot, add 1-2 tablespoons of vegetable oil … "

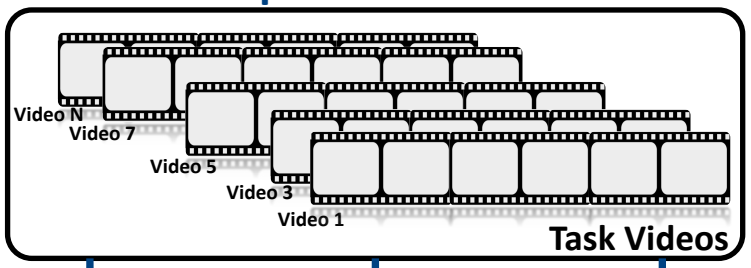
.....

Step 5
text: "Once your vegetables are tender, add your protein to the pan and stir-fry it with the vegetables …"
context: "This should take about 2-3 minutes, depending on the type and quantity of protein you are using."

.....

Step N
text: "Finally, taste and adjust the seasonings as needed."
context: "You can add more soy sauce, salt, pepper, or any other seasonings you like to suit your taste preferences."

Given Multiple Instructional Videos



V2T-B VLog

Video 1 Caption Generation . . . Video 5 Caption Generation . . . Video N Caption Generation

Prompt: According to the captions above, then prepare the logical video captions for the task procedures of the [TASK] step-by-step in a sentence template of <person> <verb> <action>.

LLM

Fusion of Captioning

Step 1
video_caption: "A person is pouring oil on a pan …
task_procedures: "Preparing Ingredients:"
.....
Step 5:
video_caption: "A person stirring a pan of food with…"
task_procedures: "Cooking Process:"
.....
Step N:
video_caption: "A bowl of food with an egg on top of it."
task_procedures: "Presentation and Enjoyment:"

Visually Grounded Textual Plan = Textual Instruction + Video Caption Procedure

Prompt: Rewrite the textual instruction of [TASK] with knowledge from visualized instruction pair-wisely in a template <text> <context>, <visual> separately.

Textual Output

Revised Text Procedural Plan

Step 1
text: "Preparing Ingredients"
context: "2 cups cooked rice (preferably day-old rice), 1 cup kimchi, … "
visual: "A person is holding a knife and cutting kimchi and onions into small pieces"
.....
Step 5
text: "Adding Cooked Rice"
context: "A bowl of cooked rice is shown on the side."
visual: "The person is adding the cooked rice to the pan and stirring it ..."
.....
Step N
text: "Seasoning and Serving"
context: "Kimchi fried rice is served by adding seaweed seasoning"
visual: "The person is sprinkling soy sauce over the kimchi fried rice and placing it on a plate"

T2V-B ModelScope

Video Output

Text-to-Video Generation