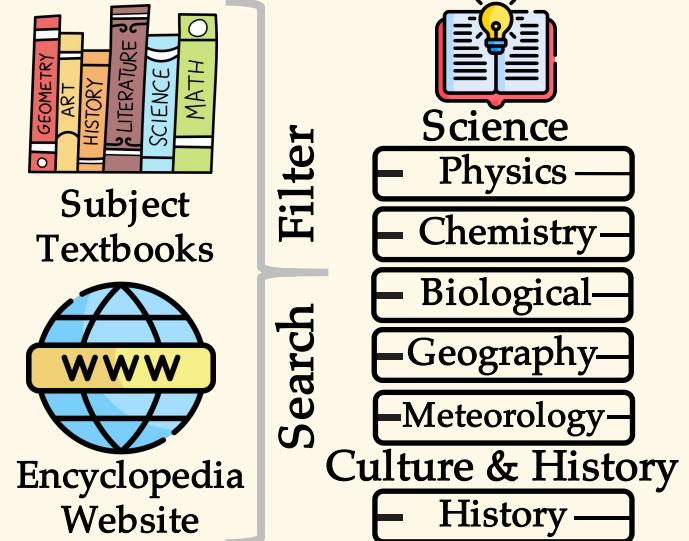


Dataset Suite



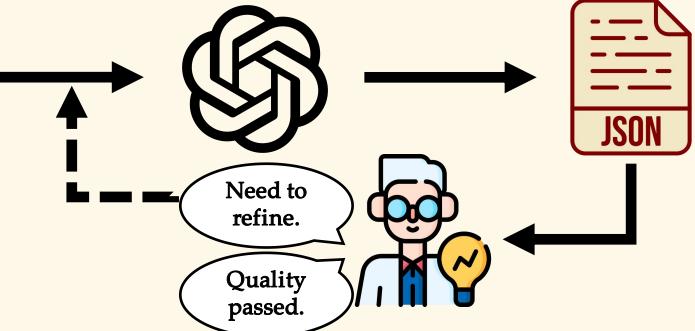
Prompt Template

```
"step": 1/2/3/4,  
"prompt": "This is the xth  
event frame of a continuous  
four-stage event sequence  
showing...",  
"explanation": "This step  
depicts ..."
```

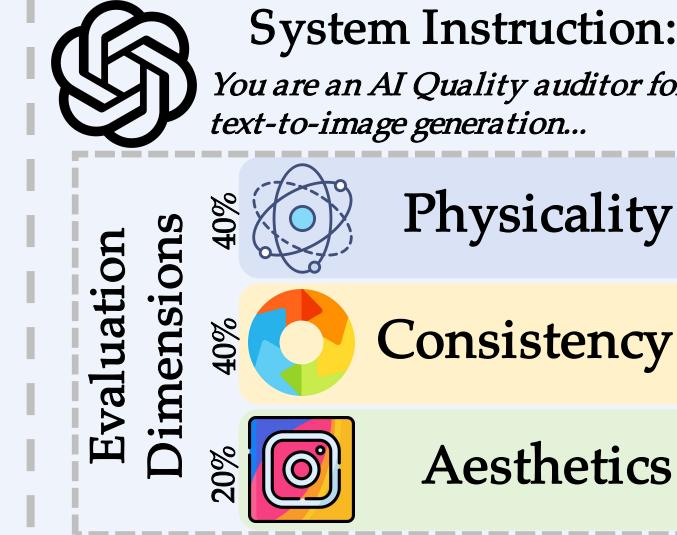
Dimension

The Prompt Generation interface features a large text input field labeled 'Instruction' containing the placeholder text: "You're a ... Clearly define...".

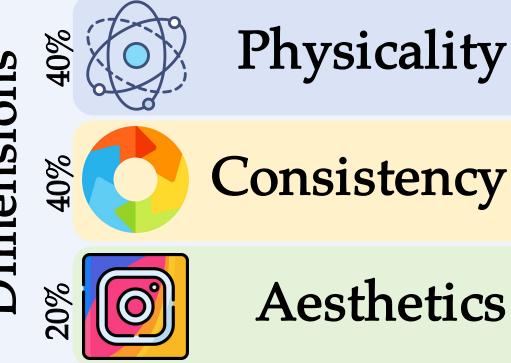
Structure



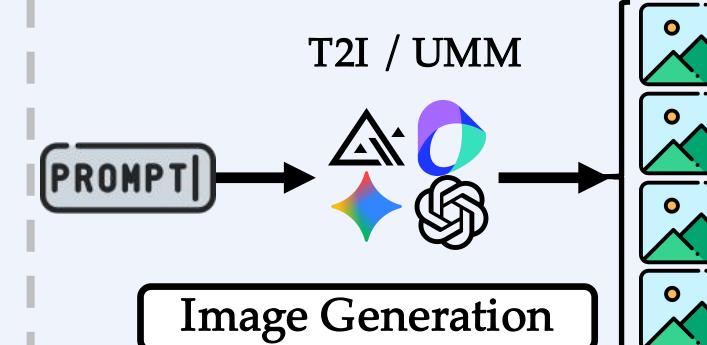
Evaluation Suite



Evaluation Dimensions



VLM Setting



Metric Analysis

Robustness and stability



Human evaluate relevance



Metrics attribute mining



Scoring Criteria	
(0-5)	
0 (Failure):	3 (Fair):
1 (Very Poor):	4 (Good):
2 (Poor):	5 (Excellent):

VLM as Judge