ETVA: Evaluation of Text-to-Video Alignment via Fine-grained Question Generation and Answering

Kaisi Guan [1,2*] Zhengfeng Lai[2] Yuchong Sun[1] Peng Zhang[2]
Wei Liu[2] Kieran Liu[2] Meng Cao[2] Ruihua Song[1†]
[1] Renmin University of China [2]Apple
https://eftv-eval.github.io/etva-eval

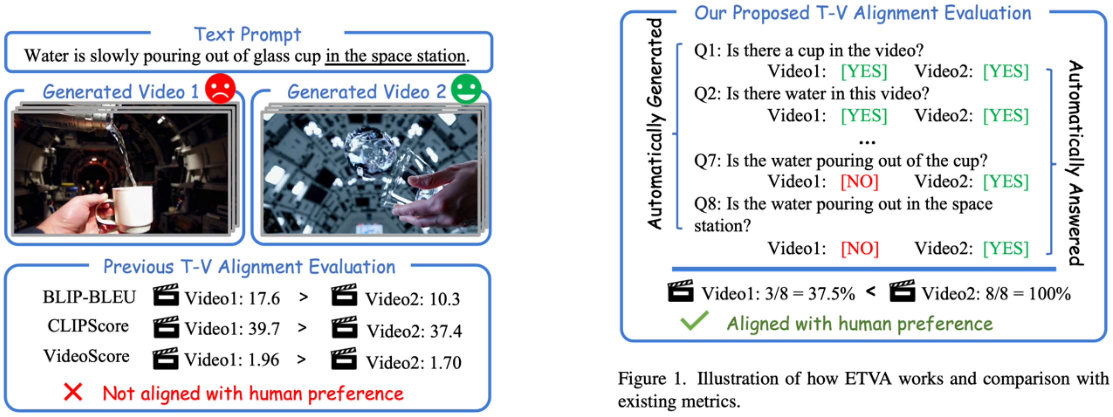
Text Prompt
Water is slowly pouring out of glass cup in the space station.
Generated Video 1
Generated Video 2
Previous T-V Alignment Evaluation
BLIP-BLEU Video1: 17.6 > Video2: 10.3
CLIPScore Video1: 39.7 > Video2: 37.4
VideoScore Video1: 1.96 > Video2: 1.70
✕ Not aligned with human preference
Our Proposed T-V Alignment Evaluation
Automatically Generated
Automatically Answered
Q1: Is there a cup in the video?
Video1: [YES] Video2: [YES]
Q2: Is there water in this video?
Video1: [YES] Video2: [YES]
…
Q7: Is the water pouring out of the cup?
Video1: [NO] Video2: [YES]
Q8: Is the water pouring out in the space station?
Video1: [NO] Video2: [YES]
Video1: 3/8 = 37.5% < Video2: 8/8 = 100%
✓ Aligned with human preference

Figure 1. Illustration of how ETVA works and comparison with existing metrics.

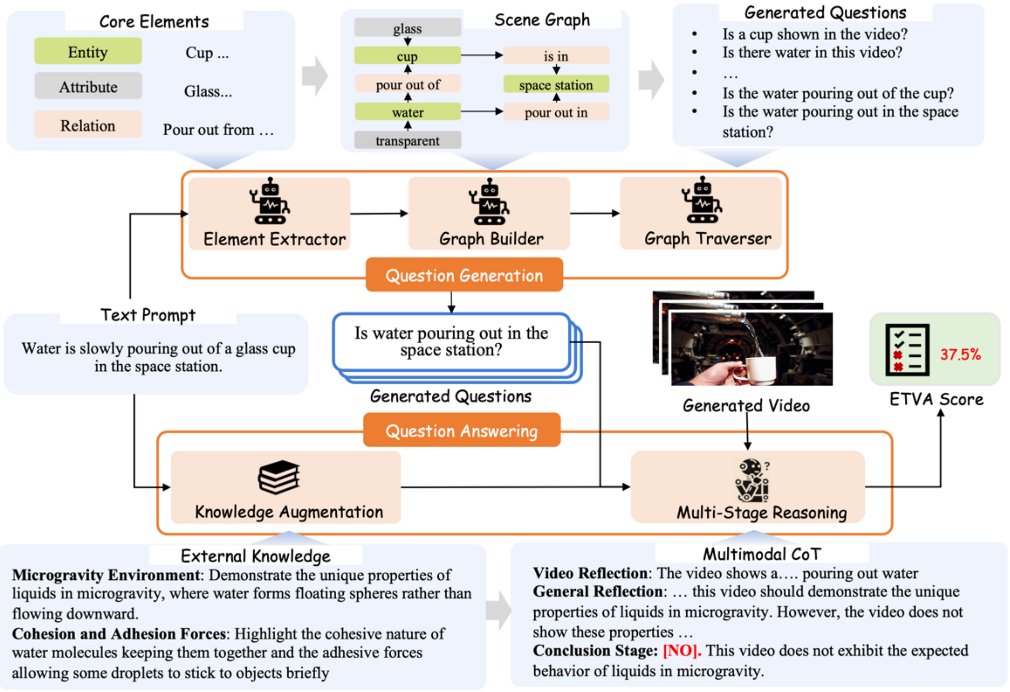
Core Elements
Entity
Cup ...
Attribute
Glass...
Relation
Pour out from …
Scene Graph
glass
cup
is in
pour out of
space station
water
pour out in
transparent
Generated Questions
Is a cup shown in the video?
Is there water in this video?
…
Is the water pouring out of the cup?
Is the water pouring out in the space station?
Element Extractor
Graph Builder
Graph Traverser
Question Generation
Text Prompt
Water is slowly pouring out of a glass cup in the space station.
Is water pouring out in the space station?
Generated Questions
Generated Video
37.5%
ETVA Score
Question Answering
Knowledge Augmentation
Multi-Stage Reasoning
External Knowledge
Microgravity Environment: Demonstrate the unique properties of liquids in microgravity, where water forms floating spheres rather than flowing downward.
Cohesion and Adhesion Forces: Highlight the cohesive nature of water molecules keeping them together and the adhesive forces allowing some droplets to stick to objects briefly
Multimodal CoT
Video Reflection: The video shows a…. pouring out water
General Reflection: … this video should demonstrate the unique properties of liquids in microgravity. However, the video does not show these properties …
Conclusion Stage: [NO]. This video does not exhibit the expected behavior of liquids in microgravity.

Figure 2. Overall pipeline of ETVA. ETVA contains a multi-agent framework for generating atomic questions and a knowledge-augmented multi-stage reasoning framework for question answering.