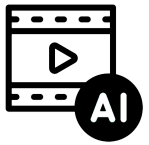

Stage 1
Physics-Grounded Question Generation

Upsampled Video Prompts
(from Video Generation Pipeline)

Enriched with Physical Details and Causal Clarity

Generate Yes/No Questions

- Aligned with 7 Physical Commonsense Dimensions
- Generated using Upsampled Prompts

Ground truth answer is **Always** Yes.

Answerable **SOLELY** by inspecting the upsampled prompts
(no external knowledge needed)

Output
Physics-Grounded Questions Targeting Different Dimensions

Stage 2
Dense Video Captioning

AuroraCap
(Video Captioning Model)

Generate 8 different dense captions per each video

Captioning Strategy
1 x General-Purpose Caption
7 x Dimension-Targeted Captions

Commonsense Dimensions

- Object Properties & Affordances
- Spatial Reasoning
- Temporal Dynamics
- Action & Procedural Understanding
- Material Interaction & Transformation
- Force and Motion

Stage 3
LLM as a Judge

Targeted Yes/No Physics Questions
(from Stage 1)

Ask Questions
Ask targeted physics questions based on the generated captions.

Scoring Logic
If LLM output is Yes (at least for one of the 8 captions), count as correct

Reliable and Interpretable Physical Commonsense Score

More reliable than traditional VQA or direct VLM scoring

Final Model Accuracy
Compute proportion of correctly answered questions