# CinePile: A Long Video Question Answering Dataset and Benchmark

Ruchit Rawal[1] Khalid Saifullah[1] Ronen Basri[2] David Jacobs[1] Gowthami Somepalli*[1] Tom Goldstein*[1]

[1]University of Maryland, College Park [2] Weizmann Institute of Science

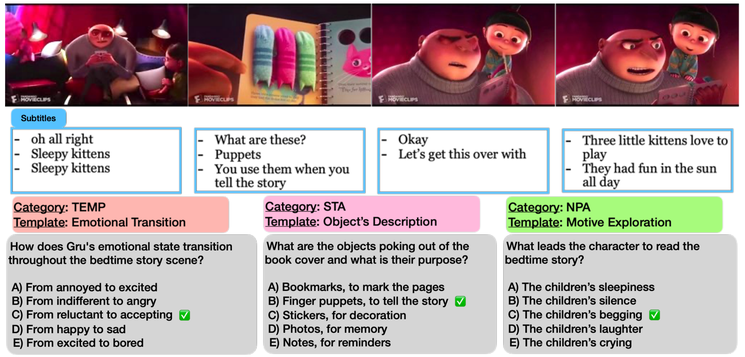


Figure 1. A sample clip (Despicable Me - Bedtime Story) and corresponding MCQs.

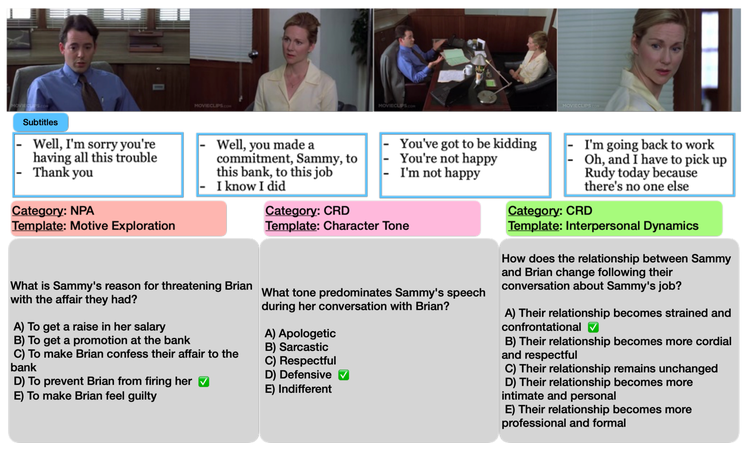


Figure 2. A sample clip (You Can Count on Me - Your Future at the Bank) and corresponding MCQs.

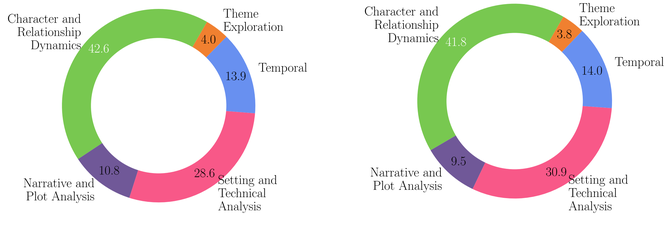


Figure 3. CinePile's question-category statistics. **Left:** Train-split. **Right:** Test-split

## Automatically Generating Question Templates

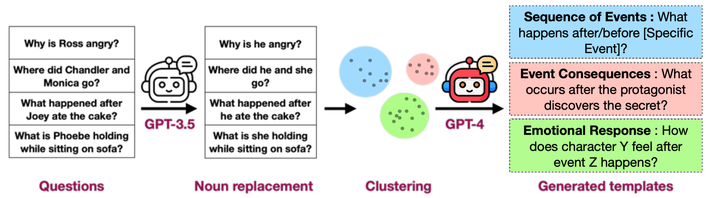


Figure 4. Question template generation pipeline: We begin by substituting the first names in human-written source questions and then cluster them. We then feed a selection of questions from each cluster into GPT-4, which outputs "question templates" used in the next stage of dataset creation.

## Automated QA Generation

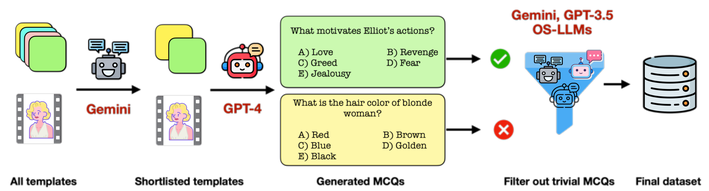


Figure 5. Automated QA Generation and Filtering. Begins with a set of automated templates and scenes. Filter out the templates relevant to each scene. Next, pass these templates along with the annotated-scene-text to GPT-4, which is then used to create multiple-choice questions (MCQs). The generated MCQs are then subjected to numerous filters to curate the final dataset.

## Validation and Filtration Checks

To prune malformed QAs, we evaluate CinePile with the help of a few LLMs on the following axes

1. **Degeneracy.** Is the answer to the question implicit in the question itself, e.g., `What is the color of the pink house?`. They constitute a minor proportion (4.5%), and we discard these from the test-split.
2. **Vision Reliance.** Denotes whether the question requires visual descriptions to be answered correctly. Question-category wise distribution presented below.
3. **Hardness.** Gauges the difficulty of questions for the models, even when provided with full context. Manually verified by the authors.

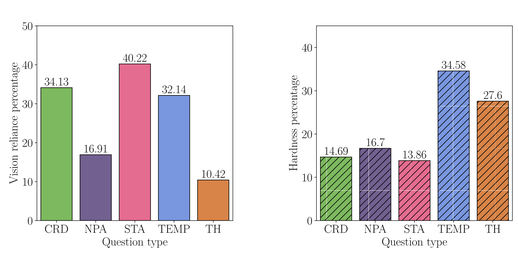


Figure 6. Distribution of Vision Reliance and Hardness across different question categories.

## Benchmarking Video-LLMs

### Performance across question categories

Table 1. **Model Evaluations.** We present the accuracy of various video LLMs on the CinePile's test split. We also present Human performance for comparison. We also ablate the accuracies across the question categories we discussed earlier. TEMP - Temporal, CRD - Character and Relationship Dynamics, NPA - Narrative and Plot Analysis, STA - Setting and Technical Analysis, TH - Thematic Exploration.

| Model | Average | CRD | NPA | STA | TEMP | TH |
|---|---|---|---|---|---|---|
| Human | 73.21 | 82.92 | 75.00 | **73.00** | 75.52 | 64.93 |
| Human (authors) | **86.00** | **92.00** | **87.5** | 71.20 | **100** | **75.00** |
| GPT-4o | 59.72 | 64.36 | **74.08** | 54.77 | 44.91 | **67.89** |
| GPT-4 Vision | 58.81 | 63.73 | 73.43 | 52.55 | 46.22 | 65.79 |
| Gemini 1.5 Pro | **61.36** | **65.17** | 71.01 | **59.57** | **46.75** | 63.27 |
| Gemini 1.5 Flash | 57.52 | 61.91 | 69.15 | 54.86 | 41.34 | 61.22 |
| Gemini Pro Vision | 50.64 | 54.16 | 65.50 | 46.97 | 35.80 | 58.82 |
| Claude 3 (Opus) | 45.60 | 48.89 | 57.88 | 40.73 | 37.65 | 47.89 |
| Video LLaVa | 22.51 | 23.11 | 25.92 | 20.69 | 22.38 | 22.63 |
| mPLUG-Owl | 10.57 | 10.65 | 11.04 | 9.18 | 11.89 | 15.05 |
| Video-ChatGPT | 14.55 | 16.02 | 14.83 | 15.54 | 6.88 | 18.86 |
| MovieChat | 4.61 | 4.95 | 4.29 | 5.23 | 2.48 | 4.21 |

- Gemini 1.5 Pro ranks as the best-performing model, performing particularly well in the "Setting and Technical Analysis" category that is dominated by visually reliant questions.
- While GPT-4 family models follow closely behind, they outperform Gemini on question categories such as "Narrative and Plot Analysis" that revolve around the core storylines.

### Performance on "hard-split"

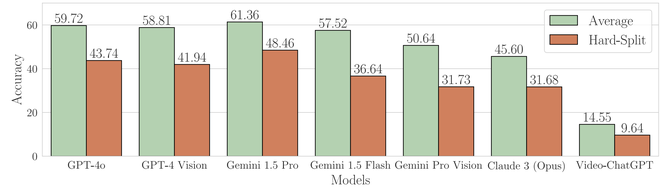


Figure 7. Models' performance on CinePile test split, all questions vs hard questions.

- Most models suffer a performance decline of 15%-20% on the hard-split.
- Gemini 1.5 Flash declines by 21% compared to 13% for Gemini 1.5 Pro, highlighting severe trade-offs in optimizing lightweight performance for challenging samples.

### Open-source Failure Modes

Table 2. Analyzing failure modes for a sample clip (Area 51 - Sneaking Onto the Base Scene).

| Question | Model Response | Failure Mode |
|---|---|---|
| How does the emotional tone transition during the scene?<br>- A) From despair to hope<br>- B) From fear to acceptance<br>- C) From confusion to understanding<br>- D) From tension to panic<br>- E) From anxiety to excitement | `"5, 4, 3, 2, 1.\n All right, I think I got it, ….. This is the base."` | The model doesn't choose any of the options, rather just repeats all the subtitles. |
| What strategy does the group employ to avoid detection?<br>- A) They communicate with the helicopter<br>- B) They use a thermal suit<br>- C) They use camouflage uniforms<br>- D) They repair a broken sensor<br>- E) They use a snake as a distraction | `The group uses a thermal suit to blend in with the surroundings and avoid detection by the helicopter.` | Ideally, the assigned score should have been 1. But we are unable to match the rephrased text. |