Narrative Aligned Long Form Video Question Answering

Supplementary Material

8. Dataset

8.1. Reasoning Dimension

We generate Question–Answer–Evidence triplets capturing 7 reasoning types detailed below.

Causal Reasoning Causal reasoning refers to the ability to identify cause–effect relationships between events that are separated in time. A model must determine why an outcome occurred by grounding its explanation in earlier observable actions, conditions, or interactions that directly lead to the effect.

Narrative Reasoning Narrative reasoning involves understanding how events unfold over time and how the story transitions from one state to another. The model must track progressions, shifts, and turning points, using a sequence of scenes that collectively show the evolution from an initial condition to a later one.

Character-Centric Reasoning Character reasoning focuses on interpreting a character's actions, decisions, and observable motivations. The model must analyze behaviors, expressions, and repeated patterns to explain why a character acts a certain way.

Thematic Reasoning Thematic reasoning captures a model's ability to infer higher-level ideas or motifs that recur throughout the story. The model must identify repeated visual patterns across multiple scenes and link them to a coherent theme.

Goal-Based Goal-Based reasoning evaluates whether the model can understand a character's goals, plans, and strategies. It requires identifying a goal (explicit or implied) and the sequence of actions taken to pursue or achieve it.

Social Reasoning Social reasoning reflects understanding of interpersonal relationships and how past interactions shape present reactions. The model must ground answers in demonstrated dynamics such as trust, conflict, loyalty, or dominance—across multiple scenes. Valid evidence includes tone of interaction, supportive or hostile behavior, shared history, and reciprocal actions.

Hypothetical Reasoning Hypothetical reasoning involves evaluating alternative possibilities grounded in the constraints of the story world. The model must compare the actual outcome with a feasible alternative scenario, using what is visually established about objects, environment, and abilities. Valid evidence includes physical limitations, shown rules, or available options.

8.2. Prompt for Dataset Construction

To obtain the initial Question–Answer–Evidence triplets, we prompt the LLM with structured event descriptions from the movie, together with clear definitions of all reasoning types. We also supply exemplars for each category to anchor the model's understanding and ensure consistent reasoning behavior. This setup helps the LLM frame questions that are grounded in observable events rather than external knowledge. The complete prompting template is provided below. The detailed prompts used for both the validator and refiner stages which are responsible for improving the quality of the dataset are also included.

9. Experiment Section

9.1. Metric

LLM as a judge metric NA-VQA is evaluated along four dimensions: Comprehensiveness, Depth, Evidence Quality, and Reasoning. For the first three dimensions, we adopt the standardized evaluation metrics introduced in MovieCORE [10], enabling consistent comparison with prior long-video benchmarks. For the Reasoning dimension, we design a dedicated LLM-based evaluation prompt that assesses whether the model demonstrates the intended causal, narrative, or character-centric reasoning. The full prompt is provided below.

Prompt for the Reasoning Metric

You are an AI evaluator designed to assess the coherence and clarity of answers to video-based questions. Your task is to evaluate whether the predicted answer is well-structured and evaluate how well the predicted answer aligns with the ground truth in terms of reasoning quality.

INSTRUCTIONS:

Judge if the predicted answer's explanation is logical, consistent, grounded, and free of hallucinations.

Scoring
Assign scores using the rubric below.
Reasoning Type Alignment – Does the answer reflect the correct reasoning type required by the question?

0: Completely misaligned reasoning approach.
1: Severely misaligned reasoning
2: Partially aligned reasoning with major gaps
3: Mostly aligned reasoning with some gaps
4: Well-aligned reasoning with minimal gaps.
5: Perfect reasoning alignment

9.2. Traditional Metrics

In addition to using LLM-as-a-judge metrics to evaluate both reasoning quality and evidence grounding, we also report results on traditional metrics commonly used in prior work. Specifically, we include BLEU [25], CIDEr, METEOR [3], and BERTScoreF1 [40] scores. These metrics capture surface-level text similarity and measure how closely a model's answer matches human-written references across n-gram overlap, sentence structure, and semantic alignment. While they do not fully reflect deeper narrative reasoning, they offer a

Table 5. Performance comparison of different models on additional metrics.

Model	Metrics			
	BL-2	MET	CID	BERT
PROPRIETARY MODELS				
Claude Sonnet 4.5	0.0403	0.2372	0.0438	0.8435
GPT-120B OSS	0.0403	0.2372	0.0438	0.8435
Qwen-243B	0.0215	0.1707	0.0372	0.8345
ZERO-SHOT VLMS				
LLaVA-OneVision (7B)	0.0098	0.1048	0.0039	0.8559
VideoLLaVA (7B)	0.0228	0.1369	0.0220	0.8188
Video-LLaMA3 (7B)	0.0361	0.1721	0.0384	0.8498
InternVL2 (7B)	0.0345	0.1742	0.0310	0.8374
InternVL2.5 (7B)	0.0387	0.1927	0.0448	0.8350
Qwen2.5-VL (7B)	0.0343	0.1747	0.0327	0.8480
Video-NaRA	0.0354	0.1781	0.0398	0.8501
ZERO-SHOT LONG-FORM VLMS				
MovieChat (7B)	0.0284	0.1527	0.0253	0.8388
VideoChat-Flash (7B)	0.0167	0.1236	0.0032	0.8314
FINE-TUNED MODELS				
Qwen2.5-VL (FT)	0.0338	0.1718	0.0244	0.8481
Video-NaRA (FT)	0.0366	0.1833	0.0458	0.8499

standardized comparison point with existing long-video QA benchmarks. Across all of these metrics, our findings remain consistent Table 5. Video-NaRA (FT) outperformed Qwen2.5 VL (FT) across all evaluated metrics. Notably, Video-NaRA achieved the highest score on the CIDEr metric.

9.3. Narrative Memory Construction

Our model, Video-NaRA, takes both the clip descriptions and the raw frames to construct the narrative memory. Specifically, for each clip, we get the detail description that captures what is happening and who/what is involved. Description along with visual clip are passed to an MLLM that creates narrative slots based on shared entities, temporal continuity, and high-level scene semantics, allowing related events to be grouped even when they appear far apart in the movie. This gives us a compact yet expressive memory bank that reflects how the story unfolds and preserves the key evidence needed for retrieval. A detailed prompt for narrative-memory construction is below.

Prompt for Narrative Memory Construction

You are an expert visual reasoning model organizing a movie into {N} fixed narrative slots.
Each slot represents one coherent storyline that unfolds across multiple clips.
A narrative slot is defined by:
Character Continuity – the same people, animals, or key objects reappear visually.
Goal or Action Continuity – the ongoing intention or action sequence remains consistent. (for example, a person running away, two people arguing, or a dog being rescued.)
Plot/Sequence of Events – the logical cause and effect.
You are shown representative frames and description from each existing slot, followed by video from a new clip. Your task is to decide which slot the new clip most likely continues.
Follow this reasoning process:
- Compare the new clip's visual content to each slot's frames.
- Look for shared characters, repeated actions, and similar emotional tone.
- Choose the slot that represents a continuation of the same story or situation.
- If no slot clearly matches (new characters, new goal, different emotion), assign it to the first unused slot (lowest ID without clips).
- Always output only one slot ID.

Return your decision **strictly in JSON format**, with no code fences, no extra text:
{ "slot": <int >, "reason": "¡brief explanation of your reasoning¿" }
Example valid outputs:
{ "slot": 0, "reason": "Same woman and dog continue walking together" }
{ "slot": 3, "reason": "New character and setting, different tone" }

Prompt for Refiner

You are an expert **Video QA Question–Answer–Evidence Refiner**. You will receive:
1. A set of **movie events descriptions**,
2. An **original Question–Answer–Evidence triplet**, and
3. **feedback** indicating scores of criteria and explanation.

Your task:
Refine and correct the Question, Answer, and Evidence so they fully satisfy **all** criteria below while strictly adhering to the visual content described in the events.
Grounding
- Only use information explicitly shown or visually inferable from the provided events.
- Do NOT add events, motivations, or facts not present in the events.

Use Validation Feedback
- For each criterion marked with score 0, fix the issue in the refined QA-Evidence.
- If required evidence is missing in the events, adjust the question/answer accordingly. Do NOT invent new events.

Quality Requirements
- **Video-Grounded**: The question must be answerable by watching the events alone.
- **Answer Faithfulness**: The answer must be strictly supported by the chosen evidence.
- **Events Completeness**: Include only the minimal events necessary; remove irrelevant ones.
- **Clarity & Challenge**: The question must be clear, unambiguous, and non-trivial.
- **Reasoning Required**: The question must require narrative, causal, theme, reasoning, etc.
- **Character-Agnostic**: Do not use character names. Use visual descriptions only (e.g., "the woman in the blue dress," "the older man," etc.).
- **Content-Identifiability**: Everything must be visually interpretable without relying on dialogue unless described in the scenes.

Evidence must directly support the answer.

Prompt for Validator

You are an expert video QA validation and correction assistant.
Your sole task is to critically evaluate a given Question, Answer, Evidence triplet based on a set of events from a movie.
You must follow these guidelines strictly:

Evaluation Criteria:
Please answer each question with 0, 1 and 2. Give a short, relevant explanation.
- **Video-Grounded Framing**: Is the question framed so it can be answered by watching the video? (No external knowledge.)
- **Answer Faithfulness**: Is the answer factually correct **based only** on the provided events?
- **Events Completeness**: Are all the necessary events present to answer the question?
- **Minimal Events**: Are there any irrelevant events that do not add to the answer?
- **Clarity & Challenge**: Is the question unambiguous and non-trivial (i.e., does it require interpretation, not surface detection)?
- **Reasoning Required**: Does the question require reasoning (narrative, causal, theme, etc.), not just recall?
- **Character-Agnostic**: Do the question, answer and evidences **avoid** using character names and instead describe observable content?
- **Content-Identifiability**: Can the question and answer be understood using **only the visual content** (not relying on dialogue)?

Prompt for Initial Question Answer Evidence Triplet

You are a rigorous teacher in a graduate-level course on Long-Term Video Understanding.
Your task is to create challenging, short-answer comprehension questions from a set of detailed scene-level events from a movie.

Each question must:
- Focus on a **single event or small group of events**, without relying on external knowledge.
- Require **deep reasoning**, not surface-level recall.
- Be **grounded in the video content**: the correct answer must be inferable **only from the provided scenes**.
- Include the corresponding **evidence**: which scene(s) (with indices) are required to answer the question.
- Be **specific** to one of the reasoning types and distance(context) categories listed below.

Question Reasoning Types and Some Examples:

1. **Causal Reasoning**: Asks "Why did X happen?" Here is the definition {Causal Definition} Requires connecting a later event (effect) to an earlier event or condition (cause) across multiple scenes. Must rely on observable cause-effect chains.

2. **Narrative Reasoning**: Asks "How did the story progress from A to B?" Here is the definition {Narrative Definition} Requires tracking changes over time. Must use a sequence of scenes showing transitions or turning points (before $\rightarrow$ after).

3. **Character Reasoning**: Asks "Why did the character do X?" Here is the definition {Character Definition}. Requires grounding in observable behavior, prior actions, dialogue, or expression. Must not assume unshown internal motivations (e.g., "he felt sad" unless shown via crying or withdrawal).

4. **Thematic Reasoning**: Asks "What does this scene reveal about a recurring idea?" Here is the definition {Thematic Definition} Requires at least 2–3 scenes that show a repeated pattern (e.g., alone in frame, closed blinds, untouched meals).

5. **Goal-Based Reasoning**: Asks "How was X achieved?" Here is the definition {Goal-Based Definition} Requires identifying a goal (stated or implied) and a sequence of actions to achieve it. Must show goal $\rightarrow$ obstacles $\rightarrow$ tools $\rightarrow$ solutions.

6. **Social Reasoning**: Asks "Why did A react to B that way?" Here is the definition {Social Definition} Requires evidence of shared history (past conflict, trust, loyalty) shown across scenes. Must link current behavior to **earlier interactions**.

7. **Hypothetical Reasoning**: Asks "What if X hadn't happened?" or "Could they have done Y?" Here is the definition {Hypothetical Definition} Requires comparing the actual outcome with a feasible alternative based on video constraints. Must ground the answer in established rules or limits.

Generate questions for each distance level if possible:

The questions should have evidence and distance between evidence should vary based on the criteria below. Each question should have evidences between 2 to 20.

SHORT Distance: Multi-hop questions that connect information across adjacent 2-4 scenes.

MEDIUM Distance: Multi-hop questions that connect information across multiple scenes at some distance apart 10-15 scenes distance.

FAR Distance: Multi-hop questions that connect information across multiple scenes at some distance apart 20-40 scenes distance.

Now, generate challenging answers question pair from the provided list of events.

For each question:
- Clearly state the question.
- Provide a 3 to 4 lines of **accurate answer** grounded in the events. Make sure events are sufficient and necessary.
- Mention the **reasoning type** and the **scene distance type**.
- Remove the nouns from the questions and answer and make sure question and answer are in such a way that it can be answered by looking at the video.

Each item in JSON list format:

```
[
  {
    "question": "...",
    "answer": "...",
    "evidence_events": [event, event, event......],
    "reasoning_type": "causal",
    "scene_distance": "medium"
  },
  ...
]
```

Here is the event information {event_info}

Figure 5. Prompt template for generating question-answer-evidence triplets with reasoning type and scene distance annotations.

Question: How does the man achieve the goal of creating a convincing family portrait despite the unusual circumstances surrounding the child?

Events:
1 A man retrieves a camera with tripod from storage to capture a family photograph.
2 The woman holds the blonde toddler wrapped in a yellow blanket on the couch while the man sets up the camera.
3 The man anxiously sets up a camera in the kitchen for a family photograph.
4 The man gives animated instructions to the woman and the toddler about the photograph.
5 The man positions himself near the woman and the baby with the camera's cable release in hand to complete their portrayal as a proper family unit.
6 The man grins at the camera while seated with the woman and the baby for what appears to be a family portrait.

Answer: The man achieves the goal of creating a convincing family portrait through careful staging and props. He retrieves a camera with tripod from storage, demonstrating planning and intention to document their family appearance. He gives animated instructions about positioning and expressions while setting up the shot. The woman holds the blonde toddler wrapped in a yellow blanket on their floral-patterned couch, creating a traditional family tableau. Despite their nervous body language revealing their discomfort with the deception, they successfully create visual evidence of their pretend normal family life.

a

Question: Why did the store employee place the handwritten sign on the metal security shutters?

Events:
1 A store employee arrives at the grocery store to find it locked behind security shutters.
2 The employee struggles with a jammed padlock due to their gloved hands.
3 The employee lies on the convenience store floor to create a makeshift sign.
4 The employee stands precariously on a garbage can outside the store.
5 The employee attaches a handwritten sign reading 'I ASSURE YOU WE'RE OPEN!' to the metal security shutters.
6 A handwritten sign reading 'I ASSURE YOU WE'RE OPEN!' hangs on the storefront.

Answer: The store employee placed the handwritten sign reading 'I ASSURE YOU WE'RE OPEN!' because they found the entrance locked behind security shutters with a jammed padlock. After struggling with the lock while wearing gloves and showing growing frustration, they improvised this solution to signal to potential customers that business was continuing despite the closed appearance of the storefront.

b

Figure 6. Additional samples from our NA-VQA dataset.