Appendix

A. Key Notation

Table 3. Key notation used throughout the paper

Symbol	Description
$X = \{x_t\}$	Continuous multimodal data stream
$\mathcal{S}_t$	Segmentation trigger indicator at time t
$\mathbb{1}[\cdot]$	Indicator function (returns 1 if true, 0 otherwise)
m_i	i-th `ShortTermMemory` object
K	Set of indices for consolidated memories
$k \in K$	Index in the consolidated set
m_k	Consolidated `ShortTermMemory` object
θ_k	k-th ThetaEvent object
$\mathcal{M}$	Collection of consolidated `ShortTermMemory` objects
Θ	Collection of ThetaEvent objects
$\mathbf{v}_i, \mathbf{v}_k$	Representative embedding for segment i or k
$\mathbf{E}_i$	Sequence of cross-modal embeddings for segment i
$\mathbf{T}_i$	Aggregated textual outputs for segment i
$\mathbf{C}_i$	Pointers to raw audiovisual data for segment i
$\mathbf{S}_{\theta_k}$	Semantic summary of ThetaEvent k
$\mathcal{E}_c$	Cross-modal embedding model
$\mathcal{T}_v$	Vision-language model for visual descriptions
$\mathcal{T}_a$	Speech recognition model
ϕ_{LLM}	LLM for semantic summary generation
γ	Consolidation similarity threshold
τ	Confidence threshold for retrieval switching
δ	Temporal window expansion parameter
$\cos(\cdot, \cdot)$	Cosine similarity function
$\text{sim}(\cdot, \cdot)$	Similarity function (cosine similarity)
$\text{TopK}(\cdot, k)$	Function returning k highest-scoring elements
$\text{FormattedContext}(\cdot)$	Function to structure multimodal content
$\text{Extract}_{\text{target}}(\cdot)$	Function to extract target modality content
$(t_{s,i}, t_{e,i})$	Start and end times of segment i
$[t_{s,i}, t_{e,i}]$	Alternative notation for segment boundaries
Φ_{fast}	Fast retrieval function
Ψ_{detailed}	Detailed recall function
Γ	Adaptive reasoning module
$\mathcal{C}_{\text{context}}$	Context information for reasoning

B. HippoVlog Construction Details

The HippoVlog dataset was created through a systematic pipeline designed to capture rich multimodal information from daily vlogs and generate challenging memory-based questions. The process follows five main stages: video selection, multimodal data sampling, ground truth generation, question generation, and curation.

B.1. Video Selection and Preprocessing

- **Source Selection:** 25 daily vlogs from YouTube (Creative Commons licensed), with average duration of 27.3 minutes, totaling approximately 682 minutes of content.
- **Quality Criteria:** Videos were chosen based on clear audio and visual quality, natural episodic structure typical of daily life, and diverse activities (cooking, shopping, travel).
- **Format Standardization:** All source videos were preprocessed to a standard format (720p resolution at 30fps) with normalized audio levels.

B.2. Multimodal Data Sampling

To create manageable and representative data points from the full vlogs, we implemented a robust sampling strategy for both audio and video modalities:

- **Audio Segment Extraction:**
  - 10 non-overlapping audio segments (each 10 seconds long) were randomly selected per video.
  - Sampling avoided the initial and final 5% of videos to exclude typical intros/outros.
  - Silent segments were discarded and resampled.
- **Video Segment Extraction:**
  - Corresponding 10-second video segments were extracted for each valid audio segment.
  - Checks prevented significant overlap between sampled segments.
  - Metadata about the sampled segments was preserved for future reference.

B.3. Segment-Level Ground Truth Generation

Detailed descriptions combining audio and visual information were generated for each sampled 10-second video segment:

- **Model Used:** Qwen2.5-Omni-7B.
- **Process:** The model processed each video segment to generate comprehensive captions.
- **Structured Output:** For each segment, a structured caption containing:
  - Detailed audiovisual description
  - List of distinct visual elements
  - Transcription of speech or description of other audio elements
  - All content maintained anonymity (using "the person", "the speaker")

B.4. Question Generation

Multiple-choice questions were generated across four categories, with visual questions specifically based on individual frames extracted from video segments:

- **Models Used:** GPT-4o.
- **Input:** For visual questions, the model received frames and corresponding video context.
- **Design Principles:** Questions were designed to:
  - Focus on specific, distinctive details
  - Be potentially answerable if the video had been previously seen
  - Maintain anonymity
  - Have exactly one correct answer among four plausible options

B.5. Question Curation and Finalization

The automatically generated questions underwent further curation:

- **Initial Pool:** Approximately 1,500 candidate questions generated across all types.
- **Quality Control:**
  - Manual validation by human.
  - Questions that were ambiguous, subjective, or had multiple plausible answers were removed or refined.
- **Final Dataset:** 1,000 high-quality questions, balanced across four categories (250 per category).
- **Answer Generation:** Each question includes one correct answer and three plausible but verifiably incorrect distractors, balanced for length and style.

B.6. Question Categories and Examples

Each of the four question categories in HippoVlog targets a specific memory function:

B.6.1. Cross-Modal Binding ($T_{V\times A}$)

Questions requiring integration of information across visual and auditory modalities.

Example: Cross-Modal Binding

Question: "When the main character is sitting in the car wearing a black jacket and white scarf, what does she audibly express?"
Options:

- A: She says she feels excited about the drive.
- B: She mentions being scared and says it feels weird.
- C: She talks about the weather being too cold.
- D: She expresses confidence about her driving skills.

Explanation: In the video, as the main character sits in the car wearing a black jacket and white scarf, she audibly expresses her nervousness by saying 'okay, hey, I'm scared, oh my gosh this feels weird.'

B.6.2. Auditory-Focused Retrieval (T_A)

Questions focusing on auditory information, including speech content, background sounds, and music.

Example: Auditory-Focused Retrieval

Question: "What did the main character mention about the purpose of the meeting?"
Options:

- A: To finalize a business deal
- B: To discuss the upcoming project
- C: To plan a surprise party
- D: To talk about a vacation

Explanation: The main character specifically mentions the 'upcoming project' as the purpose of the meeting, indicating a discussion centered around future plans and tasks.

B.6.3. Visual-Focused Retrieval (T_V)

Questions targeting visual details that require careful observation of scene elements.

Example: Visual-Focused Retrieval

Question: "What distinctive pattern is visible on the roof of the building?"
Options:

- A: Checkerboard
- B: Chevron
- C: Striped
- D: Polka dot

Explanation: The roof of the building in the image has a distinctive chevron pattern.

Example: Visual-Focused Retrieval

Question: "What is the main character wearing on her head?"
Options:

- A: A baseball cap
- B: A knitted hat
- C: A colorful headscarf
- D: A headband

Explanation: The main character is wearing a colorful headscarf, as seen in the image.

B.6.4. Semantic/Temporal Reasoning (T_S)

Questions requiring integration of information across temporal segments and abstract understanding.

Example: Semantic/Temporal Reasoning

Question: "What was one significant challenge the main character faced during the experiment?"
Options:

- A: Finding time to eat breakfast
- B: Dealing with brain fog
- C: Avoiding social media
- D: Lack of exercise options

Explanation: Later in the vlog, after the experiment has been underway for several hours, the main character states they are "dealing with a lot of brain fog," which is affecting their focus on the task.

C. HippoMM Implementation Details

This appendix provides the specific implementation choices, models, and parameters used in the HippoMM system, complementing the conceptual description in the Methodology section.

System Configuration All experiments reported were conducted on a single NVIDIA L40S GPU.

C.1. Memory Formation Implementation

Temporal Pattern Separation The process segments continuous video streams based on perceptual changes:

- **Input Sampling:** Raw video frames are adaptively sampled at a rate between 1 and 10 frames per second (fps). Raw audio is sampled at 16 kHz PCM.
- **Visual Change Detection:** A boundary is triggered ($\mathcal{S}_t = 1$) if the visual dissimilarity between consecutive frames, measured as $d_v(F_t, F_{t-1}) = 1 - \text{SSIM}(F_t, F_{t-1})$ (Eq. 2), exceeds a threshold $\tau_v = 0.65$.
- **Auditory Change Detection:** A boundary is triggered ($\mathcal{S}_t = 1$) if the negative log energy of the raw audio signal, $d_a(a_t)$ (Eq. 3), exceeds a threshold $\tau_a = 40$. This corresponds to the signal amplitude dropping below a threshold (approx. -40 dB), indicating potential silence or a significant break.
- **Segment Duration:** The resulting segments are constrained to be between $t_{min} = 5$ seconds and $t_{max} = 10$ seconds long. Shorter segments following boundary detection are merged with preceding ones, and longer intervals between boundaries are split to adhere to this constraint, defining the temporal windows $[t_{start,i}, t_{end,i}]$.

Perceptual Encoding Each segment i (defined by $[t_{start,i}, t_{end,i}]$) is processed to extract multimodal features:

- **Cross-Modal Embeddings:** We use ImageBind (v1.0) as the joint embedding model ($\mathcal{E}_c$) to generate sequences of shared 1024-dimensional embeddings ($\mathbf{E}_i$) from visual frames and audio snippets within the segment.
- **Audio Transcription:** Whisper (v3 medium) is employed as the speech recognition model ($\mathcal{T}_a$) to obtain time-aligned transcriptions from the segment's audio.
- **Visual Description:** Qwen2.5-VL is used as the vision-language model ($\mathcal{T}_v$) to generate textual descriptions of the visual content (maximum length 200 tokens).
- **`ShortTermMemory` Object (m_i):** These processed elements are aggregated and stored in a `ShortTermMemory` object m_i, as defined in Eq. (4). It encapsulates: pointers to the raw audiovisual snippets ($\mathbf{C}_i$), the sequence of ImageBind embeddings ($\mathbf{E}_i$), the aggregated Whisper transcriptions and Qwen-VL visual descriptions ($\mathbf{T}_i$), segment start/end timestamps ($t_{s,i}, t_{e,i}$), and related metadata.

Memory Consolidation Redundancy is reduced by filtering similar consecutive segments:

- **Similarity Calculation:** A representative cross-modal embedding ($\mathbf{v}_i$) is calculated for each segment m_i by averaging its sequence of ImageBind embeddings ($\mathbf{E}_i$).

- **Filtering Criterion:** A segment m_i is retained in the consolidated set K only if the cosine similarity between its representative embedding $\mathbf{v}_i$ and the representative embedding $\mathbf{v}_j$ of the most recent previously retained segment m_j (where $j \in K, j < i$) is below a predefined threshold $\gamma = 0.85$, as per Eq. 5. Segments not meeting this criterion are discarded (or merged). The output is the sequence of consolidated objects $\{m_k \mid k \in K\}$.

Semantic Replay Abstracted representations are generated for long-term storage:

- **Summarization Model:** The Qwen2.5-VL large vision-language model (ϕ_{LLM}) processes formatted multimodal context derived from the consolidated segment m_k (including key visual information $\mathbf{I}_{\text{visual}}(m_k)$ and auditory information $\mathbf{A}_{\text{audio}}(m_k)$). It generates a concise textual summary ($\mathbf{S}_{\theta_k}$) using a temperature of 0.7 and a maximum length of 300 tokens (see Table 5 for prompt details), as per Eq. 6.
- **`ThetaEvent` Object (θ_k):** This object constitutes the long-term memory entry, storing the elements defined in Eq. 7: the representative cross-modal embedding ($\mathbf{v}_k$), the generated textual summary ($\mathbf{S}_{\theta_k}$), references to the key visual ($\mathbf{I}_{\theta_k}$) and auditory ($\mathbf{A}_{\theta_k}$) context used, and the segment's temporal information ($\mathbf{T}_{\theta_k}$) which links back to the detailed `ShortTermMemory` object m_k. The collection $\{\theta_k\}$ forms the abstract long-term memory store.

C.2. Memory Retrieval Implementation

Query Analysis User queries are initially processed to determine the retrieval strategy:

- **Classifier Model:** GPT-4o is used as the query classifier ($\mathcal{Q}_{\text{type}}$) to classify the incoming query q based on its primary modality focus (Visual, Auditory, Cross-modal) or if it targets high-level semantic/gist information (Summary). This uses a 3-shot prompt with examples (see Table 6).

Hierarchical Retrieval Retrieval follows a confidence-gated two-pathway approach (Eq. 8):

- **Fast Retrieval (Φ_{fast}):** Retrieval is first attempted using the abstract textual summaries ($\mathbf{S}_{\theta_k}$) stored in the collection of `ThetaEvent` objects ($\mathbf{\Theta} = \{\theta_k\}$). Confidence scores are determined by internal LLM calibration (Qwen2.5-VL, see Table 7) on a scale of 0-1.
- **Detailed Recall (Ψ_{detailed}):** If the confidence score from Fast Retrieval falls below a threshold $\tau = 0.75$, retrieval escalates to the Detailed Recall pathway. This pathway searches the consolidated `ShortTermMemory` objects ($\mathcal{M} = \{m_k \mid k \in K\}$), guided by the query type determined by $\mathcal{Q}_{\text{type}}(q)$.

Detailed Recall Pathways The specific operations within Detailed Recall (Ψ_{detailed}) are guided by the classified query type:

- **Query Embedding:** The query q is converted into an ImageBind embedding (q_{embed}) for use in similarity-based search.
- **Similarity Search (Feature/Semantic):** For single-modality or cue-based retrieval (Visual or Auditory query type), cosine similarity is used to find the top-k `ShortTermMemory` segments (m_k) whose representative embeddings ($\mathbf{v}_k$) are most similar to q_{embed} (Feature Search), with $k = 5$. Text-based search over the aggregated descriptions/transcriptions ($\mathbf{T}_k$) serves as a fallback or alternative (Semantic Search). See Tables 8 and 9 for related selection prompts.
- **Cross-Modal Association:** For Cross-modal query types, retrieval uses temporal co-occurrence. First, the top-k seed segments $\mathbf{S}_{\text{query}}$ most similar to q_{embed} (based on $\mathbf{v}_k$) are identified (Eq. 9, with $k = 5$). Second, expanded temporal windows ($\mathbf{W}$) are defined around these seeds using a temporal buffer parameter $\delta = 2$ seconds (Eq. 10). Third, target modality information ($\mathbf{S}_{\text{target}}$) is extracted from any segment m_j whose time interval overlaps with any window in $\mathbf{W}$ (Eq. 11).
- **Fine-grained Analysis:** If necessary for answering the query, specific content within the retrieved `ShortTermMemory` segments is analyzed further using Qwen2.5-VL (for visual/multimodal questions) or Whisper (for detailed audio content examination). This often occurs during the final synthesis stage.

Adaptive Reasoning The final answer is synthesized from the retrieved information:

- **Synthesis Model:** Qwen2.5-VL is typically used as the reasoning module (Γ) to generate the final answer (a). It takes the original query (q), the retrieved evidence ($r_{\text{retrieved}}$), and relevant extracted context ($\mathcal{E}_{\text{context}}$) as input (Eq. 12, see Table 11 for prompt). For complex reasoning tasks requiring deeper integration or handling potential conflicts (e.g., comparing Fast Retrieval and Detailed Recall outputs), GPT-4o may be used with a 4-shot prompting strategy and temperature of 0.2 (see Table 10 for a related prompt).

Table 4. Cross-benchmark generalization comparison. Video-MME and LongVideoBench test general comprehension; MLVU-NQA tests precise temporal retrieval, where HippoMM's episodic memory excels. Best results in **bold**.

Method	Video-MME	LongVideoBench	MLVU-NQA
InternVL2 [4]	61.2%	59.3%	48.3%
VideoLLaMA 3 [41]	**66.2%**	**59.8%**	68.3%
GPT-4o [15]	71.9%	66.7%	64.8%
HippoMM (Ours)	61.8%	55.2%	**73.1%**

D. Cross-Benchmark Generalization

E. Additional Experimental Details

E.1. Ablation Studies Parameter Settings

Acronyms refer to components described in the Methodology section: FR (Fast Retrieval, Φ_{fast}), DR (Detailed Recall, Ψ_{detailed}), AR (Adaptive Reasoning, Γ).

- **HippoMM w/o DR, AR:** Uses Fast Retrieval (Φ_{fast}) only, with direct answer generation from summaries (no Adaptive Reasoning Γ).
- **HippoMM w/o FR, AR:** Uses Detailed Recall (Ψ_{detailed}) only, with direct answer generation from retrieved details (no Adaptive Reasoning Γ).
- **HippoMM w/o AR:** Uses both retrieval paths (Φ_{fast}, Ψ_{detailed}) as per the hierarchical logic, but the final output is based on direct extraction or simple aggregation rather than the full synthesis step (Γ).

E.2. Performance Metrics

- **Processing Time (PT):** Total time to execute the Memory Formation phase on raw videos.
- **Average Response Time (ART):** Average time to execute the Memory Retrieval phase and generate an answer a for a query q.
- **Accuracy Metrics:** Based on strict binary match with human-annotated ground truth answers.

E.3. Generalization Evaluation

- **Charades Dataset:** 100 randomly selected video clips.
- **MLVU Needle-in-a-Haystack:** 415 queries across 150 long-form videos (5+ minutes each).
- **GPT-4o Evaluation:** Used as judge for Content Relevance, Semantic Similarity, and Action Alignment with a 5-point Likert scale.
- **Overall Score Calculation:** Arithmetic mean after linearly normalizing 1-5 scales to 0-100%.

E.4. Additional Implementation Parameters

- **Key Frame Filtering Threshold:** Within Perceptual Encoding, a visual similarity threshold (SSIM-based, value 0.3) is used to filter visually redundant frames before description generation to reduce processing load. (Note: This is distinct from the consolidation threshold γ).
- **Visual Similarity Fallback Threshold (Detailed Recall):** When performing Feature Search in Detailed Recall for visual queries, if the max similarity score is below 0.4, the system may fall back to Semantic Search using text captions.
- **Direct Answer Confidence Threshold (τ):** The threshold used in Hierarchical Retrieval (Eq. 8) to switch from Fast Retrieval to Detailed Recall is $\tau = 0.75$.
- **Context Length (LLM):** A maximum context length of 120,000 tokens is used for GPT-4o when invoked for complex reasoning (Adaptive Reasoning Γ).
- **Multiple Worker Processing:** Multiprocessing with a pool size equal to the number of available CPU cores (up to a maximum of 8) is used for parallel frame processing during Perceptual Encoding.

Table 5. Semantic Replay Summarization Prompt (ϕ_{LLM})

Task Description Please provide a concise one sentence summary of this event based on the video frames descriptions and audio transcription. What is happening in this event?

Input Format

- Image descriptions: [Concatenated frame captions from key visual context $\mathbf{I}_{\text{visual}}(m_k)$]
- Audio transcription: [Concatenated transcribed text from key auditory context $\mathbf{A}_{\text{audio}}(m_k)$]

Output Requirements

- One sentence summary ($\mathbf{S}_{\theta_k}$) capturing the integrated multimodal essence of the event
- Focus on key actions, subjects, and setting
- Maintain factual accuracy based on provided descriptions
- Avoid introducing information not supported by the input

Table 6. Query Type Classification Prompt ($\mathcal{Q}_{\text{type}}$)

Task Description Classify this question into one of these categories based on what type of information is needed to answer it:

Classification Categories

- VIDEO - Questions specifically about visual elements, appearances, or actions that need frame-by-frame analysis, e.g., "what is the main character holding?" (Corresponds to Visual V)
- AUDIO - Questions about sounds, speech, or audio content that need audio analysis, e.g., "what does the main character say/mention?" (Corresponds to Auditory A)
- VIDEO+AUDIO - Questions that require associating both visual and audio information, e.g., "what is the main character doing while saying/mentioning something?" (Corresponds to Cross-modal $V \times A$)
- SUMMARY - Questions that focus on the overall content or gist of the video, e.g., "what is the main character doing?" (Corresponds to Semantic/Gist)

Guidelines

- If question is about visual details, appearances, or actions, classify as VIDEO
- If question is about sounds, speech, or audio content, classify as AUDIO
- If question requires both visual and audio information, classify as VIDEO+AUDIO
- If question is about the overall content of the video, classify as SUMMARY

Input Question: [user query q]

Output Format Return ONLY one of these exact words: VIDEO, AUDIO, VIDEO+AUDIO, SUMMARY

Table 7. Fast Retrieval Attempt Prompt (Part of Φ_{fast})

Task Description Given the following question and relevant event summaries ($\mathbf{S}_{\theta_k}$ from $\mathbf{\Theta}$), analyze whether the question can be answered directly or needs specific analysis (triggering Detailed Recall Ψ_{detailed}).
Input Question: [user query q]
Event Summaries: [List of relevant event summaries $\mathbf{S}_{\theta_k}$ retrieved based on q]
Guidelines for analysis
- General questions about overall video content (likely classified as SUMMARY) should be attempted directly from summaries.
- Questions about specific visual details (VIDEO), sounds/speech (AUDIO), or combined elements (VIDEO+AUDIO) often require specific analysis beyond summaries.
- Estimate confidence based on whether summaries contain sufficient information.

Output Format Must be in one of these two structures:
1. If answerable from summaries:

```
ANSWER: <your detailed answer based ONLY on summaries,
if multiple choice, output one letter>
CONFIDENCE: <score between 0.0-1.0, e.g., 0.7 or higher if confident>
```

2. If requiring specific analysis (summaries insufficient):

```
ANSWER: NONE
CONFIDENCE: <low score, e.g., 0.0 or < 0.7>
```

(Note: This confidence score is compared against threshold τ)

Table 8. Visual Segment Selection Prompt (Part of Ψ_{detailed} - Semantic Search)

Task Description Given a question (potentially classified as VIDEO or VIDEO+AUDIO), descriptions from consolidated memory segments ($\mathbf{T}_k$), and an optional element to search for, identify at most 5 relevant frame indices (within segments m_k) for answering the question.
Input Question: [user query q] Element to search for: [search query derived from q]
Frame descriptions (from relevant $\mathbf{T}_k$): [Numbered list of frame descriptions/captions]
Instructions
- Return ONLY numbers separated by commas (e.g., "0,3,5,8,12") corresponding to indices within the provided list.
- Return at most 5 indices.
- Do not include any other text, explanations, or spaces.
- If fewer than 5 frames are relevant, return fewer indices.

Example good responses "0,3,5,8,12" "1,4,7"
Output Format Your response (numbers only, comma-separated):

Table 9. Audio Segment Selection Prompt (Part of Ψ_{detailed} - Semantic Search)

Task Description Given a question (potentially classified as AUDIO or VIDEO+AUDIO) and transcriptions with timestamps from consolidated memory segments ($\mathbf{T}_k$), identify the most relevant time frames where the answer might be found.
Input Question: [user query q]
Transcriptions (with timestamps from relevant $\mathbf{T}_k$): [List of timestamped transcriptions]
Instructions
- Analyze the transcriptions and identify segments most likely to contain the answer.
- Return a JSON array of time frames (relative to the source video). Each frame defines a start and end time.

```
[
    {"start": START_TIME, "end": END_TIME},
    {"start": START_TIME, "end": END_TIME}
]
```

- Return at most 5 time frames.
- Include a small buffer around identified time frames (e.g., ± 2 seconds, related to δ).
- If no relevant segments found, return "[]".

Output Format Your response (valid JSON only):

Table 10. Adaptive Reasoning Prompt (Optional step within Γ)

Task Description You potentially have two candidate answers to the same question q. One derived from Fast Retrieval (Φ_{fast}, using summaries $\mathbf{\Theta}$) and another from Detailed Recall (Ψ_{detailed}, using detailed segments $\mathcal{M}$). Compare them and provide a final, reconciled answer a.
Input Question: [user query q]
Answer from summaries (if Φ_{fast} produced one): [direct_answer] Confidence from summaries: [direct_confidence]
Answer from detailed analysis (if Ψ_{detailed} was run): [detailed_answer]
Supporting Context ($\mathcal{E}_{\text{context}}$): Sample frame captions: [Sample relevant frame captions from $r_{\text{retrieved}}$] Sample transcriptions: [Sample relevant transcriptions from $r_{\text{retrieved}}$]
Instructions
- Compare both answers for consistency, detail, and plausibility given the context.
- If they agree, prefer the more detailed/confident answer.
- If they disagree, evaluate which answer is better supported by the specific context (captions, transcriptions). Prioritize the answer derived from detailed analysis (Ψ_{detailed}) if confidence in the summary-based answer (Φ_{fast}) was low or if the detailed analysis provides clearly contradictory evidence.
- Synthesize the best possible final answer based on this comparison.
- For multiple choice questions, select only one final answer option.
- Estimate the confidence in the final reconciled answer.

Output Format

```
ANSWER: <reconciled final answer a>
CONFIDENCE: <final confidence score between 0.0-1.0>
REASONING: <brief explanation for the final answer choice>
```

Table 11. Final Answer Synthesis Prompt (Γ)

Task Description Based on the following retrieved context ($r_{\text{retrieved}}$) and extracted evidence ($\mathcal{E}_{\text{context}}$), please synthesize a final answer (a) to the question (q).
Input Question: [user query q]
Overall Video Context (e.g., relevant summaries $\mathbf{S}_{\theta_k}$ if available): [List of relevant high-level context]
Relevant Retrieved Content (details from Φ_{fast} or Ψ_{detailed}): [List of retrieved content descriptions, e.g., frame captions, transcriptions, details from m_k]
Instructions Please provide a clear and specific answer based on the integration of the overall context and the specific retrieved content. If the information is insufficient or contradictory, state that or provide the best possible reasoned answer based on the evidence.
Output should be one letter if the original question was multiple choice.
Output Format

```
Answer: <Synthesized answer a>
```
