DIAGNOSING AND ADVANCING LONG-VIDEO UNDERSTANDING FOR DISTRIBUTED EVIDENCE COMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models can process increasingly long videos, but answering questions that depend on multiple temporally distributed observations remains challenging. We investigate how evidence demand, composition structure, and visual-context allocation affect long-video understanding. We introduce DECBench, a diagnostic benchmark comprising 926 videos and 1,123 open-ended questions spanning aggregation, temporal composition, entity-state continuity, and video-grounded explanatory reasoning and summarization. To quantify per-instance evidence demand, we introduce Minimum Evidence Count (MEC), which estimates the size of a minimum sufficient evidence set and enables comparisons across composition requirements at comparable evidence-demand levels. Our diagnostic analyses reveal two complementary bottlenecks: models often fail to acquire the necessary distributed evidence, while substantial errors persist even when relevant evidence is directly provided, indicating additional difficulty in composing the available observations. We further identify a task-dependent trade-off between temporal coverage and per-observation visual fidelity. Motivated by these findings, we propose SEEK, a training-free, query-conditioned framework that selects complementary observations according to query-implied evidence requirements and adaptively allocates visual capacity under a fixed budget. Experiments across three video MLLMs and three benchmarks show consistent improvements over query-aware frame-selection baselines. Together, DECBench and SEEK provide tools for diagnosing distributed-evidence failures and improving evidence acquisition in long-video understanding.

1 INTRODUCTION

Recent multimodal large language models (MLLMs) can process increasingly long videos (Chen et al., 2025b; Wu et al., 2024; Fu et al., 2025; Zhou et al., 2025; Wang et al., 2025a), yet long-context capacity does not guarantee long-horizon understanding. Many video questions require integrating temporally distributed observations, such as counting events, recovering temporal order, tracking entity states, or forming global explanations. These tasks require not only acquiring multiple evidence units but also composing them according to task-specific relations. While evidence demand characterizes the amount of required information, it does not capture how evidence units interact. This motivates our central question: *after controlling for evidence demand, do different relational structures of distributed evidence still pose different challenges to video understanding?*

Existing long-video benchmarks evaluate models from perspectives including video duration, task coverage, and multi-evidence dependency. Closest to our setting, HERBench (Ben Ami et al., 2026) studies multi-evidence integration with the Minimum Required Frame-Set (MRFS) and retrieval/fusion analyses. However, evidence-demand measurements alone cannot distinguish whether performance gaps arise from evidence quantity or the relations among evidence units. We therefore introduce **DECBench** and formulate **Distributed Evidence Composition (DEC)** as combining complementary visual evidence from temporally separated regions through task-dependent relations. As illustrated in Figure 1, DEC requires both acquiring distributed evidence and composing it according to question-implied structures.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

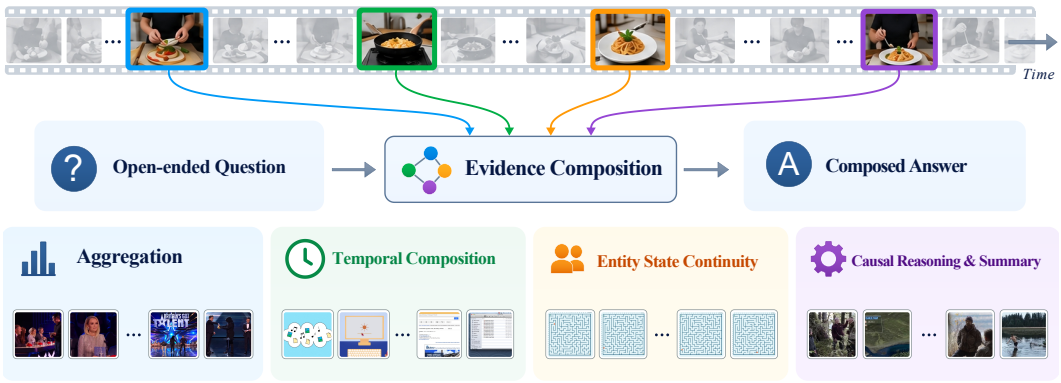


Figure 1: **Overview of DECBench.** Given an open-ended question, a model must acquire complementary evidence from temporally separated parts of a long video and compose it into an answer. DECBench captures four composition requirements: Aggregation (AG), Temporal Composition (TC), Entity State Continuity (EC), and Causal Reasoning & Summary (CS).

We organize DEC into four composition families: **Aggregation (AG)**, **Temporal Composition (TC)**, **Entity State Continuity (EC)**, and **Causal Reasoning & Summary (CS)**. By considering per-instance evidence demand and composition type, DECBench enables controlled comparisons of model behavior across different composition requirements under comparable evidence demands.

DECBench contains 926 videos and 1,123 open-ended questions, with an average video duration of 788 seconds. It combines three complementary construction routes: curation from existing benchmarks, annotation-guided construction over real videos, and programmatic video generation with complete event–state trajectories. Real videos provide diverse visual and semantic contexts, while programmatic environments allow controlled variation in evidence spacing, distractors, entity count, and state complexity. All samples are designed to require multiple temporally distributed observations whose information must be meaningfully composed.

To quantify per-instance evidence demand, we further introduce **Minimum Evidence Count (MEC)**, defined as the size of a minimum sufficient evidence set. Because exhaustive subset enumeration is impractical, we estimate MEC through evidence-requirement decomposition, coverage-guided subset construction, and sufficiency verification. MEC provides a per-instance measure of evidential demand, which can be used to compare different composition requirements while controlling for the estimated amount of required evidence.

Our analysis examines how evidence demand and visual-context construction affect performance. Matched-MEC comparisons show that composition difficulty cannot be explained by evidence quantity alone. Oracle-evidence experiments reveal two bottlenecks: acquiring the necessary distributed evidence and correctly composing it once observed. Finally, fixed-budget experiments expose a task-dependent coverage–fidelity trade-off: broader temporal coverage is preferable, whereas fine-grained entity-state reasoning benefits from greater per-frame detail.

Motivated by these findings, we propose **SEEK**, a training-free framework for **Structured Evidence Extraction with Keyframes**. SEEK maps a question into evidence slots and relations, and jointly selects complementary observations from a high-recall candidate bank. Under a fixed visual budget, it balances temporal coverage and per-observation fidelity without requiring benchmark labels, reference answers, MEC values, or annotated evidence.

Our main contributions are threefold:

- We introduce **DECBench**, a benchmark for distributed evidence composition covering four composition requirements over temporally distributed evidence.
- We develop an **MEC estimation protocol** and conduct evidence-demand-controlled diagnostic analyses, including matched-MEC comparisons, oracle-evidence evaluation, and fixed-budget allocation studies.
- We propose **SEEK**, a training-free, relation-aware evidence-selection framework that jointly selects complementary observations and adaptively allocates visual capacity; it consis-

tently improves over query-aware frame-selection baselines across three backbones and three benchmarks.

2 DECBENCH: DIAGNOSING DISTRIBUTED EVIDENCE COMPOSITION

DECBench evaluates *Distributed Evidence Composition* (DEC) in long-video understanding: whether a model can identify, retain, and compose multiple pieces of visual evidence that are separated over time. The benchmark is motivated by a simple but important distinction: *long video input does not necessarily imply long-range understanding*. A question associated with a long video may still be answerable from one salient moment, in which case the evaluation primarily measures local retrieval or recognition rather than global composition.

DECBench contains 926 videos and 1,123 open-ended question–answer pairs, with an average video duration of 788 seconds. As shown in Figure 2, the videos span a broad range of durations, from under 60 seconds to over 3,000 seconds. The benchmark also covers diverse content domains: game videos form the largest category (28.6%), followed by tutorial videos (13.0%), while the remaining samples are distributed across a variety of other domains.

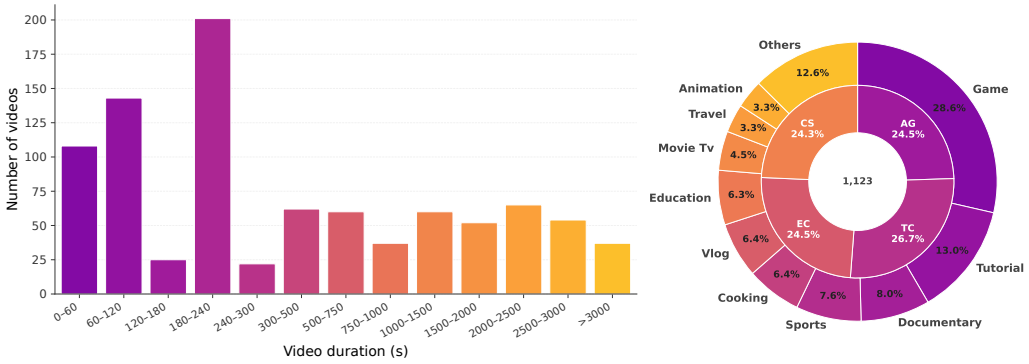


Figure 2: **Dataset statistics of DECBench.** Left: distribution of video durations across different duration intervals. Right: The inner ring shows the distribution of QA pairs across composition families, while the outer ring shows the distribution of videos across content domains.

2.1 TASK DESIGN

2.1.1 DISTRIBUTED EVIDENCE COMPOSITION

Given a video V, a question q, and a reference answer y, let $\mathcal{E}_q = \{e_1, e_2, \ldots, e_K\}$ denote a set of temporally localized visual evidence units relevant to answering q. An evidence unit can be a frame or a short temporal region carrying information needed for the answer. Let $\mathcal{E}_q^\star$ be a minimum sufficient evidence set:

$$\mathcal{E}_q^\star = \arg\min_{\mathcal{S}\subseteq\mathcal{E}_q} |\mathcal{S}| \quad \text{s.t.} \quad (\mathcal{S}, q) \Rightarrow y, \tag{1}$$

where $(\mathcal{S}, q) \Rightarrow y$ indicates that the information in $\mathcal{S}$, with the question, is sufficient to support the answer. DECBench focuses on questions for which $|\mathcal{E}_q^\star| \geq 2$ and the evidence is distributed over non-local portions of the video. We instantiate this idea through four task families.

2.1.2 TAXONOMY

Open-ended formulation. All instances use open-ended QA. Given V and q, a model directly generates an answer instead of selecting from predefined candidates. Open-ended evaluation naturally accommodates counting, ordering, state-transition, explanation, and summarization questions, while reducing shortcuts based on candidate-specific language cues or option elimination.

Aggregation (AG) requires the model to collect repeated instances, actions, events, or attributes from multiple temporal regions and combine them into a global result. Typical questions involve

total counts, frequency comparisons, or set construction. A single occurrence provides only partial information, so answering requires both broad temporal coverage and consistent accumulation.

Temporal Composition (TC) requires the model to reconstruct temporal relations among distributed events. Questions may ask for the order of several events, whether one event occurs before or after another, or how a process develops through multiple stages. Recognizing all events is insufficient unless their temporal locations are preserved and their temporal relations are resolved.

Entity State Continuity (EC) focuses on persistent entity–state binding. An entity may disappear and reappear, move between locations, change attributes, or take different roles over time. Correct answering requires maintaining identity across observations while updating the associated state.

Causal Reasoning & Summary (CS) requires higher-level semantic integration of multiple events into an explanation or global description. Questions may ask why an outcome occurs, how preceding events contribute to a later result, or what major changes occur throughout a video. Here, "causal reasoning" refers to video-grounded explanatory or narrative relations supported by observable events rather than formal causal identification under interventions.

2.2 Data Construction

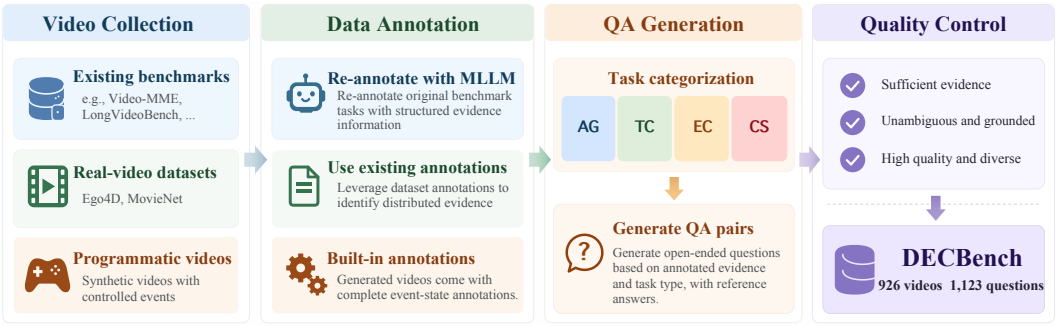


Figure 3: **Construction pipeline of DECBench.** Videos are obtained from three complementary sources and processed through four stages: data collection, data annotation, QA generation, and quality control. Source-specific annotations and simulator trajectories support open-ended QA construction, while shared filtering enforces dependency on temporally distributed visual evidence.

DECBench is constructed through a unified four-stage pipeline consisting of **video data collection, data annotation, QA generation, and quality control**, as illustrated in Figure 3. We combine three complementary sources—existing video benchmarks, annotated real-video datasets, and programmatically generated videos—to balance naturalness, structured supervision, and controllability.

Video Data Collection. We construct a candidate video pool from three sources: existing benchmarks, real-video datasets, and programmatic videos. For existing benchmarks, we curate samples from Video-MME (Fu et al., 2025), LongVideoBench (Wu et al., 2024), CG-Bench (Chen et al., 2025a), EgoSchema (Mangalam et al., 2023), and LVBench (Wang et al., 2025a), retaining videos whose questions potentially require evidence from multiple temporal regions. We additionally select real videos from Ego4D (Grauman et al., 2022) and MovieNet (Huang et al., 2020) with rich and temporally distributed actions, events, scene changes, or narratives. Samples with insufficient events, sparse annotations, homogeneous content, or mainly local cues are removed. Finally, we generate controlled videos using rule-based game environments, where duration, event density, entity count, state complexity, distractors, and evidence spacing are explicitly controlled.

Data Annotation. We collect structured evidence for QA construction. For existing benchmarks, we preserve QA information and use an MLLM to annotate temporally distributed evidence. For Ego4D and MovieNet, we leverage action-, event-, shot-, scene-, and narrative-level annotations to identify distributed evidence, supporting AG, TC, and CS. For programmatic videos, simulators provide event-state trajectories, including timestamps, entities, and state changes.

QA Generation. All samples are converted into a unified open-ended QA format. Existing benchmark questions are filtered to remove locally answerable cases, and multiple-choice questions are

converted by removing options and normalizing answers. For Ego4D and MovieNet, related annotated segments are composed to generate questions, answers, and supporting evidence. For programmatic videos, QA pairs are derived from simulator logs, where event records support AG, timestamps support TC, and entity trajectories support EC. They are not used as the main source for CS because rule-based environments tend to produce overly templated explanations.

Quality Control. We filter QA pairs to ensure that (i) at least two distinct evidence units are required, (ii) evidence participates in meaningful composition rather than co-occurrence, (iii) evidence is temporally distributed, (iv) questions cannot be solved by language priors, commonsense guessing, or multiple-choice artifacts, and (v) answers are unambiguous and video-grounded. Annotation-based samples are further reviewed for naturalness, completeness, and multi-evidence dependency, while programmatic samples are verified using complete trajectories.

2.3 Minimum Evidence Count

The definition specifies that DECBench questions require multiple evidence units, but the amount of required evidence can vary from one question to another. Video duration does not capture this property: two equally long videos may be associated with a question answerable from one salient moment and a question requiring many distributed observations, respectively. We therefore introduce **Minimum Evidence Count (MEC)** to characterize the evidential demand of a question.

Definition. For a sample (V, q, y), let $\mathcal{E}_q = \{e_1, e_2, \ldots, e_n\}$ be a candidate set of visual evidence units. We define

$$\mathrm{MEC}(q) = \min_{\mathcal{S} \subseteq \mathcal{E}} |\mathcal{S}| \quad \text{s.t.} \quad (\mathcal{S}, q) \Rightarrow y. \tag{2}$$

MEC is thus the cardinality of a minimum sufficient evidence set; larger values indicate that more visual observations must be jointly available to establish the reference answer.

Coverage-guided estimation. Exact subset enumeration is impractical, so we estimate MEC using evidence-requirement decomposition, coverage-based subset selection, and sufficiency verification. A fixed frame selector first extracts n candidate evidence frames. A fixed multimodal model then decomposes (q, y) into visual requirements $\mathcal{R} = \{r_1, r_2, \ldots, r_m\}$. For every evidence unit e_i and requirement r_j, the model estimates a support score $a_{ij} \in [0, 1]$, yielding $A = [a_{ij}] \in [0, 1]^{n \times m}$.

Let $z_i \in \{0, 1\}$ indicate whether e_i is selected. We obtain a compact candidate set by solving $\min_{\mathbf{z}} \sum_{i=1}^{n} z_i$ subject to $\sum_{i=1}^{n} a_{ij} z_i \geq 1, j = 1, \ldots, m$ and $z_i \in \{0, 1\}, i = 1, \ldots, n$.

The selected evidence is then ranked by coverage, and we search for the smallest subset from which a fixed verifier can recover an answer consistent with y. If the smallest verified set is $\widehat{\mathcal{S}}$, we report $\widehat{\mathrm{MEC}}(q) = |\widehat{\mathcal{S}}|$. Samples for which even the complete candidate set is insufficient are marked as unresolved rather than assigned an artificial evidence count.

3 Benchmark Analysis

3.1 Evaluation Setup

We evaluate 11 representative closed- and open-source MLLMs on DECBench. All models use the same questions and are instructed to answer concisely. GPT-5.6 and the open-source models use 64-frame inputs, while Gemini-3.7-Flash samples at 0.5 fps with at most 64 frames. Because DECBench uses open-ended QA, we use Qwen3-8B as an LLM judge to assign binary correctness and report accuracy overall and for each task category. Full prompts, decoding settings, and judge-validation details are deferred to Appendix B.1.

3.2 Benchmark Results

DECBench has an average MEC of 5.264, compared with 3.049 on Video-MME, 3.918 on LongVideoBench, and 4.030 on LVBench, indicating a substantially higher dependence on multiple visual evidence units (Appendix D.1).

Table 1 shows that models remain far from solving distributed-evidence composition reliably. GPT-5.6-sol reaches 54.05% overall and Gemini-3.7-Flash 44.61%, while the strongest open-source

Table 1: Performance on DECBench. Scores are accuracies (%).

Model	Input	AG	TC	EC	CS	OVERALL
GPT-5.6-sol	64	**39.64**	**62.33**	**54.18**	**59.34**	**54.05**
GPT-5.6-terra	64	30.55	51.67	41.45	49.82	43.54
GPT-5.6-luna	64	28.36	49.67	36.73	48.35	40.96
Gemini-3.7-Flash	0.5 fps / 64	32.36	56.67	38.55	49.82	44.61
Qwen3.5-27B	64	20.00	35.33	19.27	40.66	28.94
Qwen3.5-9B	64	20.36	34.67	17.45	33.70	26.71
Qwen3-VL-8B-Instruct	64	18.55	27.33	16.00	34.80	24.22
InternVL3.5-14B-Instruct	64	21.82	23.00	18.18	30.04	23.24
InternVL3.5-8B-Instruct	64	16.73	27.67	17.82	27.11	22.44
Qwen2.5-VL-7B-Instruct	64	17.45	28.00	14.18	26.01	21.55
LLaVA-OneVision-1.5-8B	64	12.73	20.67	13.45	27.47	18.61

model, Qwen3.5-27B, reaches 28.94%. These category averages can conflate composition type and evidence demand. We therefore next control for MEC before interpreting category-level difficulty.

3.3 Composition Difficulty at Matched Evidence Demand

MEC measures how many evidence units must be jointly available, but it does not specify how those units must interact. To separate evidence quantity from composition structure, we stratify samples by MEC and compare categories within the same evidence-demand level. Because MEC is integer-valued, we use the bins $\{4, 5, \geq 6\}$ and compute $\mathrm{Acc}(c \mid \mathrm{MEC} = k)$ for each category c. To ensure a balanced evaluation, we randomly sampled 40 instances from each task category under each MEC configuration. We report this controlled analysis for Qwen3.5-9B and Qwen3-VL-8B-Instruct.

Table 2: Performance under matched evidence demand.

Model	Category	MEC=4	MEC=5	MEC≥6
Qwen3.5-9B	AG	30.0	22.5	15.0
Qwen3.5-9B	TC	40.0	37.5	32.5
Qwen3.5-9B	EC	30.0	20.0	7.5
Qwen3.5-9B	CS	45.0	40.0	37.5
Qwen3-VL-8B-Instruct	AG	25.0	20.0	10.0
Qwen3-VL-8B-Instruct	TC	40.0	37.5	27.5
Qwen3-VL-8B-Instruct	EC	22.5	17.5	7.5
Qwen3-VL-8B-Instruct	CS	42.5	35.0	30.0

As shown in Table 2, substantial differences among composition categories persist after matching on MEC. EC accuracy is lower than that of TC and CS in every stratum for both models, and is no higher than that of AG. These results suggest that differences in estimated evidence demand do not fully account for the observed category gaps. MEC captures the cardinality of a minimum sufficient evidence set, whereas EC additionally requires maintaining entity identity, updating state, and associating observations across time. More broadly, these results support treating evidence demand and composition structure as complementary axes of long-video difficulty rather than collapsing them into a single notion of multi-evidence complexity.

3.4 Evidence Coverage and Oracle Inputs

To separate evidence acquisition from composition, we compare uniform sampling with an evidence-oracle condition that supplies frames containing the annotated or construction-derived evidence. Table 3 illustrates that increasing uniform samples from 32 to 64 frames yields only modest gains, whereas evidence-oracle inputs improve all three models substantially. However, performance remains limited even when the relevant evidence is made visible, indicating that evidence acquisition is not the only bottleneck and that substantial difficulty remains in interpreting and composing the available observations.

Table 3: Overall accuracy (%) under different input-frame settings.

Model	32 Uniform	64 Uniform	Oracle Frame
Qwen3.5-27B	24.58	28.94	36.69
Qwen3-VL-8B-Instruct	23.78	24.22	32.23
InternVL3.5-8B-Instruct	21.19	22.44	26.00

3.5 Fixed Visual-Budget Allocation

Finally, we investigate how to allocate a fixed visual patch budget between temporal coverage and per-observation fidelity. For Qwen3-VL-8B-Instruct, we keep the aggregate preprocessing-level patch budget fixed while varying the number of sampled frames and the per-frame patch allocation. Under the same budget, using more frames provides broader temporal coverage, while allocating more patches per frame preserves greater visual detail. These results reveal a task-dependent coverage–fidelity trade-off, motivating adaptive rather than uniform visual-budget allocation. The full category breakdown is provided in Appendix D.2.

4 SEEK: Structured Evidence Extraction with Keyframes

DECBench shows that long-video understanding depends not only on observing evidence, but also on composing observations distributed across time. We therefore propose **SEEK**, a training-free, query-conditioned framework that selects a compact evidence set under a fixed visual budget.

4.1 Query-Conditioned Evidence Planning

Given a video V, question q, and visual budget B, we densely sample a high-recall candidate bank $\mathcal{C} = \{c_i\}_{i=1}^{N}$ with timestamps τ_i. A lightweight planner outputs

$$P(q) = (\mathcal{U}, \mathcal{R}, \Theta, \rho), \tag{3}$$

where $\mathcal{U} = \{u_j\}_{j=1}^{M}$ is a set of textual evidence slots, $\mathcal{R}$ contains their required relations, $\Theta = (\lambda_e, \lambda_c, \lambda_r, \lambda_d)$ controls the selection objective, and $\rho \in [0, 1]$ controls the coverage–fidelity trade-off. We use nonnegative weights with $\lambda_e + \lambda_c + \lambda_r + \lambda_d = 1$. The planner uses only the question and requires no task label, reference answer, MEC value, or annotated evidence at inference time.

Using a frozen image–text encoder, we compute $z_i = \phi(c_i)$ and slot relevance $r_{ij} = \cos(z_i, \psi(u_j))$. Scores are rank-normalized to $\tilde{r}_{ij} \in [0, 1]$, with larger relevance mapped to larger normalized values; when $N = 1$, we set $\tilde{r}_{1j} = 1$ to avoid the zero denominator. This slot-conditioned scoring prevents one globally salient event from suppressing other required evidence.

4.2 Structured Evidence Selection

SEEK selects frames jointly. For $S \subseteq \mathcal{C}$, we maximize

$$J(S) = \lambda_e E(S) + \lambda_c C(S) + \lambda_r R(S) - \lambda_d D(S), \quad J(\emptyset) = 0 \tag{4}$$

Evidence coverage. We use

$$E(S) = \sum_{j=1}^{M} w_j \log\left(1 + \sum_{c_i \in S} \tilde{r}_{ij}\right), \tag{5}$$

where $w_j \geq 0$ and $\sum_j w_j = 1$. Hence $E(\emptyset) = 0$, while the concavity encourages coverage of multiple slots but still gives diminishing positive gain to repeated evidence.

Temporal coverage and relations. Dividing the video into H temporal bins $\{\mathcal{B}_h\}_{h=1}^{H}$, we define $C(S) = H^{-1} \sum_{h=1}^{H} \mathbf{1}[S \cap \mathcal{B}_h \neq \emptyset]$, so $C(\emptyset) = 0$. For planned relations,

$$R(S) = \sum_{(j,k,r) \in \mathcal{R}} w_{jk} \max_{\substack{c_a, c_b \in S \\ a \neq b}} \tilde{r}_{aj} \tilde{r}_{bk} K_r(\tau_a, \tau_b), \tag{6}$$

where $w_{jk} \geq 0$ and, when $\mathcal{R} \neq \emptyset$, $\sum_{(j,k,r)\in\mathcal{R}} w_{jk} = 1$. We set $R(S) = 0$ when $\mathcal{R} = \emptyset$ or $|S| < 2$, so the maximum is never taken over an empty feasible set. The constraint $a \neq b$ prevents one frame from serving as both endpoints of the *same* relation; a frame may still support multiple slots or participate in different relations. We use $K_{\text{before}} = \mathbf{1}[\tau_a < \tau_b]$, $K_{\text{after}} = \mathbf{1}[\tau_a > \tau_b]$, $K_{\text{unordered}} = 1$, and $K_{\text{progression}} = |\tau_a - \tau_b|/T$.

Redundancy suppression. To suppress adjacent near-duplicates, we define

$$D(S) = \sum_{c_i \in S} \max_{c_k \in S \setminus \{c_i\}} \text{sim}(z_i, z_k) \exp\left(-\frac{|\tau_i - \tau_k|}{\sigma}\right), \tag{7}$$

with the convention $\max \emptyset = 0$; therefore $D(\emptyset) = D(\{c_i\}) = 0$. In practice, σ is tied to the candidate sampling interval (e.g., $\sigma = 2\Delta$), so nearby duplicates are penalized more strongly than temporally separated repetitions.

Unified greedy selection. Starting from $S_0 = \emptyset$, we repeatedly select

$$c^\star = \arg\max_{c_i \in \mathcal{C} \setminus S} \Delta J(c_i \mid S), \qquad \Delta J(c_i \mid S) = J(S \cup \{c_i\}) - J(S), \tag{8}$$

and update $S \leftarrow S \cup \{c^\star\}$ until $|S| = K$. In particular, the first step is well-defined: $\Delta J(c_i \mid \emptyset) = \lambda_e E(\{c_i\}) + \lambda_c C(\{c_i\})$, because both relation and redundancy terms vanish for a singleton. Thus all questions share the same selector; query-dependent behavior enters only through the planned slots, relations, and parameters.

4.3 Adaptive Visual-Budget Allocation

SEEK uses ρ to allocate the fixed budget between temporal coverage and per-frame fidelity. The number of observations is

$$K(\rho) = \text{round}((1-\rho)K_{\max} + \rho K_{\min}), \tag{9}$$

clipped if necessary so that $1 \leq K(\rho) \leq |\mathcal{C}|$ and the resulting allocation is feasible under B. Each selected frame receives approximately $B/K(\rho)$ visual tokens. Small ρ favors broader temporal coverage, whereas large ρ favors fewer, higher-fidelity observations. Finally, selected frames are restored to chronological order and passed to the video MLLM.

Overall, SEEK converts each query into an evidence-acquisition profile and applies one shared, relation-aware selector without task-specific selection branches.

5 Experiments

We evaluate SEEK under a controlled visual-input budget, focusing on two questions: whether structured evidence acquisition improves long-video understanding across benchmarks and backbones, and which components of SEEK are responsible for the gains. We compare against representative query-aware frame-selection methods and conduct ablations on structured selection and query-adaptive evidence planning.

5.1 Experimental Setup

Benchmarks. We evaluate on DECBench, LongVideoBench, and Video-MME. DECBench targets distributed evidence composition across Aggregation (AG), Temporal Composition (TC), Entity State Continuity (EC), and Causal Reasoning & Summary (CS), while LongVideoBench and Video-MME test whether the method generalizes to established long-video benchmarks.

Backbone models. We use three open-source video MLLMs: Qwen3.5-9B, Qwen3-VL-8B-Instruct, and Qwen2.5-VL-7B-Instruct. Within each comparison, the downstream MLLM is fixed so that differences reflect the evidence-selection strategy.

Baselines and visual budget. We compare SEEK with AKS and Q-Frame, two query-aware frame-selection methods. All methods use the same 64-frame input budget, controlling for input context size. We report accuracy (%) on all benchmarks.

5.2 Main Results

Table 4 compares SEEK with AKS and Q-Frame across three benchmarks and three backbone MLLMs. SEEK achieves the best accuracy in all nine backbone–benchmark combinations. The improvement is most pronounced on DECBench for Qwen3-VL-8B and Qwen2.5-VL-7B, where SEEK reaches 27.78% and 23.95%, respectively, outperforming both baselines by more than one percentage point. This supports the central motivation of SEEK: selecting a complementary set of observations is particularly useful when answering requires evidence distributed across time.

The improvement also transfers to LongVideoBench and Video-MME for every backbone. Although the gains on these benchmarks are generally smaller than on DECBench, their consistency indicates that structured evidence acquisition is not specific to the proposed benchmark.

Table 4: Comparison of frame-selection methods on DECBench, LongVideoBench, and Video-MME. Scores are accuracies (%). Best results within each backbone are shown in bold.

Backbone	Method	DECBench	LongVideoBench	Video-MME
Qwen3.5-9B	AKS	27.52	68.46	70.70
	Q-Frame	27.78	69.21	71.07
	SEEK	**28.22**	**69.98**	**71.37**
Qwen3-VL-8B-Instruct	AKS	25.73	64.53	67.67
	Q-Frame	26.36	65.83	67.11
	SEEK	**27.78**	**66.02**	**67.81**
Qwen2.5-VL-7B-Instruct	AKS	21.91	62.45	64.52
	Q-Frame	22.08	63.36	65.78
	SEEK	**23.95**	**63.91**	**66.07**

5.3 Ablation Studies

We conduct ablation studies on DECBench with Qwen3-VL-8B-Instruct to analyze the contribution of each component in SEEK. Detailed category-level results are provided in Appendix D.3.

Structured evidence selection. We progressively add Temporal Coverage, Relation Consistency, and Redundancy Suppression to the relevance-only baseline. As shown in Table 7, each component consistently improves performance, with the full SEEK objective achieving the strongest results. The gains demonstrate that effective evidence acquisition requires jointly modeling temporal coverage, evidence relations, and redundancy rather than relying on individual frame relevance alone.

Query-adaptive evidence profile. We compare a Fixed Profile, an oracle-style GT Task-specific Profile, and Query-adaptive SEEK. Query-adaptive SEEK consistently outperforms the fixed strategy across all composition categories and substantially closes the gap with the oracle profile. Without access to task labels, reference answers, MEC values, or annotated evidence during inference, it achieves similar acquisition behavior to the task-specific oracle. This demonstrates that the question itself contains sufficient information to infer the required evidence composition pattern, enabling adaptive evidence allocation without additional supervision.

6 Conclusions

In this work, we study distributed evidence composition in long-video understanding and introduce DECBench together with Minimum Evidence Count (MEC) to characterize composition requirements and evidence demand. Our analyses reveal two complementary bottlenecks: acquiring distributed evidence and correctly composing it, as well as a task-dependent trade-off between temporal coverage and visual fidelity. Based on these findings, we propose SEEK, a training-free, query-conditioned framework for structured evidence extraction and adaptive visual-budget allocation. Experiments across multiple models and benchmarks consistently demonstrate its effectiveness. Overall, our results highlight structured evidence acquisition and composition as important directions for advancing long-video understanding.

AI USE STATEMENT

Generative AI tools were used to assist with the writing and editing of this manuscript, including improving clarity, grammar, and readability. They were also used to assist with literature retrieval and discovery, such as identifying potentially relevant prior work and related references. All retrieved literature and bibliographic information were independently verified by the authors against the original sources before inclusion in the manuscript. Generative AI tools were not used to develop the research methodology, design experiments, generate or analyze experimental data, interpret experimental results, or formulate the scientific conclusions of this work. All AI-assisted content was reviewed and revised by the authors, who take full responsibility for the final manuscript, including its citations, claims, and conclusions.

ETHICS STATEMENT

DECBench is constructed from existing research benchmarks and datasets, including Video-MME, LongVideoBench, CG-Bench, EgoSchema, LVBench, Ego4D, and MovieNet, together with programmatically generated videos. We use these resources for academic research in accordance with their applicable licenses, data-use agreements, and access terms. In particular, LongVideoBench and LVBench are distributed under CC BY-NC-SA 4.0; EgoSchema is distributed under the Ego4D license and applicable access terms; Ego4D is accessed under its data-use agreement; and Video-MME, CG-Bench, and MovieNet are used subject to their respective research-use and distribution terms. Where source datasets do not grant redistribution rights to the underlying videos, such videos remain subject to the rights of their original copyright holders and are not relicensed by us as part of DECBench.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the methods and experimental protocols needed to reproduce our results. The construction procedure and Minimum Evidence Count (MEC) estimation protocol for DECBench are described in Sections 2.2–2.3. The SEEK framework, including query-conditioned evidence planning, structured evidence selection, and adaptive visual-budget allocation, is specified in Section 4. Experimental settings and evaluation protocols are provided in Section 5 and Appendix B, with additional implementation details, analyses, and ablation studies included in the appendices. Upon completion of the conference review process, we will publicly release the DECBench annotations and associated benchmark resources, together with the implementation of SEEK. Source videos will be distributed or made accessible in accordance with the licenses and terms of their respective source datasets.

REFERENCES

Dan Ben Ami, Gabriele Serussi, Kobi Cohen, and Chaim Baskin. Herbench: A benchmark for multi-evidence integration in video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4505–4514, 2026.

Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations*, volume 2025, pp. 45647–45682, 2025a.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. In *International Conference on Learning Representations*, volume 2025, pp. 18227–18246, 2025b.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Sunqi Fan, Jiashuo Cui, Meng-Hao Guo, and Shuojin Yang. Tool-augmented spatiotemporal reasoning for streamlining video question answering task. *Advances in Neural Information Processing Systems*, 38:129243–129273, 2025.

Sunqi Fan, Qingle Liu, Runqi Yin, Meng-Hao Guo, and Shuojin Yang. Bridging videoqa and video-guided agentic tasks via generalized keyframe extraction. In *European Conference on Computer Vision (ECCV)*, 2026.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in MLLMs. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *European conference on computer vision*, pp. 709–727. Springer, 2020.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

Jihao Qiu, Lingxi Xie, Xinyue Huo, Qi Tian, and Qixiang Ye. Longvideo-r1: Smart navigation for low-cost long video understanding. *arXiv preprint arXiv:2602.20913*, 2026.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal adaptive compression for long video-language understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54582–54599, 2025.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18221–18232. IEEE, 2024.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29118–29128. IEEE, 2025.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22958–22967. IEEE, 2025a.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025b.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 3272–3283, 2025c.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, et al. Longvt: Incentivizing" thinking with long videos" via native tool calling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33816–33826, 2026.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, et al. Frame-voyager: Learning to query frames for video large language models. In *International Conference on Learning Representations*, volume 2025, pp. 84154–84179, 2025.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long Context Transfer from Language to Vision. *Transactions on Machine Learning Research*, 2025b.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22056–22065, 2025c.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025d.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701. IEEE, 2025.

A Related Work

Video MLLMs. Video MLLMs have rapidly evolved from general video instruction tuning to long-context modeling. Representative recent models include VideoLLaMA2 (Cheng et al., 2024), LongVA (Zhang et al., 2025b), LongVILA (Chen et al., 2025b), LLaVA-Video (Zhang et al., 2025d), and MovieChat (Song et al., 2024). Other systems further extend context length or reduce video-token cost, including LongVU (Shen et al., 2025), InternVideo2.5 (Wang et al., 2025b) and VideoLLaMA3 (Zhang et al., 2025a). In parallel, Video-R1 (Feng et al., 2025) and VideoChat-R1 (Li et al., 2025) explore reinforcement learning for video reasoning and spatiotemporal perception, while recent methods such as LongVideo-R1 (Qiu et al., 2026) and LongVT (Yang et al., 2026) target efficient long-video navigation and tool-assisted reasoning. Despite this progress, long-context capacity alone does not ensure that multiple temporally distant evidence units are jointly preserved and composed; DECBench focuses on this capability.

Long-Video Benchmarks. Recent benchmarks have advanced long-video evaluation across diverse tasks and durations. Video-MME (Fu et al., 2025), MLVU (Zhou et al., 2025), and LVBench (Wang et al., 2025a) focus on broad long-context understanding, while CG-Bench (Chen et al., 2025a) targets clue-grounded answering. HERBench (Ben Ami et al., 2026) is closest to our setting, studying multi-evidence integration with the Minimum Required Frame-Set (MRFS). In contrast, **DECBench** organizes distributed evidence composition into aggregation, temporal composition, entity-state continuity, and causal reasoning/summary, and uses open-ended QA with Minimum Evidence Count (MEC) to analyze evidence interactions under comparable evidence demand.

Keyframe and Evidence Selection. Efficient long-video reasoning relies on selecting informative observations under limited budgets. AKS (Tang et al., 2025) balances query relevance and temporal coverage, Q-Frame (Zhang et al., 2025c) combines query-aware selection with multi-resolution processing, while Frame-Voyager (Yu et al., 2025), VideoTree (Wang et al., 2025c), STAR (Fan et al., 2025), and TASKER (Fan et al., 2026) explore query-conditioned selection, hierarchical search, or task-aware extraction. In contrast, SEEK represents query-implied evidence requirements as slots and relations, guiding joint selection of complementary observations and adaptive visual-token allocation.

B More Details of Evaluation

B.1 Evaluation Protocol

We evaluate models on DECBench using a unified open-ended video question answering protocol. Each sample consists of a video V and an open-ended question q, and the model is required to directly generate an answer without access to candidate options. This formulation is designed to evaluate whether a model can acquire and compose temporally distributed evidence rather than exploit multiple-choice cues.

For benchmark evaluation, we use the same question set for all models and instruct models to answer concisely. Unless otherwise specified, video inputs are sampled with 64 frames. For GPT-5.6 and the open-source video MLLMs, we use uniformly sampled 64-frame inputs. For Gemini-3.7-Flash, we sample frames at 0.5 fps with a maximum of 64 frames. These settings follow the input constraints described in the main experiments.

Since DECBench contains open-ended questions, exact string matching is insufficient for evaluation. We therefore use an LLM-based judge to determine answer correctness. Specifically, Qwen3-8B is used as the evaluator and assigns a binary correctness label by comparing the generated answer with the reference answer and considering whether the response is supported by the video content. The reported metric is accuracy, computed as the percentage of questions judged correct. We report both the overall accuracy and category-level accuracy for the four composition requirements: Aggregation (AG), Temporal Composition (TC), Entity State Continuity (EC), and Causal Reasoning & Summary (CS).

To improve evaluation reliability, the judge is provided with the question, reference answer, and model-generated answer, and is instructed to focus on semantic correctness rather than exact wording. Minor differences in phrasing are accepted when the answer preserves the required evidence and composition relations. Answers containing unsupported claims or missing essential evidence are marked incorrect.

For controlled diagnostic experiments, we keep the evaluation procedure unchanged and only modify the visual input condition. In matched-MEC analysis, samples are stratified according to their estimated Minimum Evidence Count (MEC), and category comparisons are conducted under the same evidence-demand level. In oracle-evidence experiments, models receive frames containing annotated or construction-derived evidence units instead of uniformly sampled frames, allowing us to isolate evidence acquisition from evidence composition ability.

B.2 Prompt Templates

To ensure consistent evaluation across different video MLLMs, we use carefully designed prompts that explicitly define the task objective, answer requirements, and output constraints. The prompts avoid requiring models to reveal intermediate reasoning and only evaluate the final answer grounded in the video content.

For all benchmark models, we use the same instruction template. The template emphasizes evidence grounding and distributed evidence composition, while avoiding task-specific hints that could reveal the category of each sample.

Video QA Evaluation Prompt

System Instruction
You are a helpful video understanding assistant. Your task is to answer questions by analyzing the provided video.
Follow these rules:

1. Use only information that can be verified from the video content.
2. Carefully inspect relevant observations across the entire video before answering.
3. When a question requires multiple events, combine evidence from different temporal locations instead of relying on a single moment.
4. Preserve temporal order, entity identity, state changes, and causal relationships when they are required by the question.
5. Do not use external knowledge, assumptions, or common-sense guesses that are not supported by the video.
6. If the video does not contain sufficient evidence, state that the information cannot be determined from the video.
7. Provide only the final answer. Do not output analysis, intermediate reasoning, or evidence lists unless explicitly requested.
8. Keep the answer concise while including all information necessary to address the question.

User Input
Video: {video}
Question: {question}
Answer:

LLM Judge Evaluation Prompt

System Instruction
You are an expert evaluator for open-ended video question answering. Your goal is to determine whether a predicted answer correctly answers the given question according to the reference answer.
Evaluation rules:

1. Judge semantic correctness rather than exact string matching.
2. Accept different wording, synonyms, and equivalent expressions if they convey the same meaning as the reference answer.
3. Require that the answer is supported by the video and contains the essential information needed to answer the question.
4. Mark an answer incorrect if it:
   - contradicts the reference answer or video content;
   - relies on unsupported assumptions;
5. Do not penalize minor differences in wording, grammar, or answer length.
6. Do not evaluate writing quality; only evaluate factual and semantic correctness.

Output exactly one label: `Correct` or `Incorrect`.
Evaluation Input
Question: {question}
Reference Answer: {reference}
Predicted Answer: {prediction}
Evaluation:

C More Details of Data Construction

DECBench is constructed with the goal of evaluating distributed evidence composition rather than local video recognition. Therefore, each instance is required to contain multiple temporally separated evidence units whose information must be jointly composed to answer the question. The construction pipeline consists of four stages: data source collection, evidence annotation, open-ended QA generation, and quality control.

C.1 Data Sources

We collect videos from three complementary sources to balance naturalness, annotation richness, and controllability.

Existing Video Benchmarks. We collect candidate videos from existing long-video understanding benchmarks, including Video-MME, LongVideoBench, CG-Bench, EgoSchema, and LVBench. These benchmarks provide diverse video domains and existing annotations, which serve as both direct QA sources and video resources for further reconstruction.

For samples with suitable existing questions, we preserve the original question-answer pairs and convert them into our unified open-ended QA format. Meanwhile, to better align with distributed evidence composition, we also revisit the original videos and construct new QA pairs based on their visual content and temporal structures. Specifically, we identify videos containing multiple relevant events, entity changes, or long-range temporal dependencies, annotate their distributed evidence, and redesign questions that require composing information from multiple temporal regions.

This reconstruction process allows us to utilize the large-scale coverage of existing benchmarks while avoiding cases where the original questions mainly evaluate local recognition or short-range retrieval. Samples that can be answered from a single salient moment, rely heavily on multiple-choice artifacts, or lack sufficient distributed evidence are removed.

Real-world Video Datasets. To increase the diversity of temporal structures, we further incorporate videos from large-scale real-world datasets such as Ego4D and MovieNet. These datasets contain rich temporal annotations, including actions, events, shots, scenes, and narrative structures. We select videos containing multiple events, state changes, or long-range interactions that naturally support distributed evidence composition.

Programmatically Generated Videos. Although real-world videos provide realistic visual complexity, their evidence structures are difficult to precisely control. Therefore, we additionally generate synthetic videos using rule-based environments. Each generated video is associated with complete event-state trajectories, allowing explicit control over factors including evidence spacing, number of entities, event density, state transitions, and distractor events.

The three sources provide complementary advantages: real videos introduce natural complexity, existing benchmarks provide broad task coverage, and programmatic videos enable controlled analysis of evidence requirements.

C.2 Evidence Annotation

To support distributed evidence composition, every candidate sample is annotated with structured evidence information. The annotation aims to identify which temporal observations are necessary for answering a question and how they should be composed.

For existing benchmark samples, we preserve original question-answer pairs and annotate supporting evidence regions. A multimodal large language model is used to identify relevant temporal segments and associate them with the information required by the question. The generated annotations are further filtered to remove cases where the predicted evidence is incomplete or concentrated in a single local region.

For real-world video datasets, we leverage existing annotations such as action segments, event boundaries, scene transitions, and narrative descriptions. These annotations are converted into evidence units that describe temporally localized information relevant to question answering.

For programmatically generated videos, evidence annotations are directly obtained from simulator trajectories. Each trajectory records event timestamps, entity identities, state changes, and object interactions. These annotations provide exact evidence locations and enable fine-grained control over evidence composition patterns.

Each evidence unit contains: (1) a temporal location, (2) a visual description of the observed event or state, (3) the associated entity information when applicable, and (4) its role in answering the target question.

Based on these annotations, each sample is categorized into one of four composition requirements: Aggregation (AG), Temporal Composition (TC), Entity State Continuity (EC), and Causal Reasoning & Summary (CS).

C.3 QA Generation

After obtaining structured evidence annotations, we generate open-ended QA pairs through an evidence-driven generation process. Different from conventional video QA construction that first samples questions from videos, DECBench starts from the annotated evidence structure and explicitly designs questions whose answers require composing multiple evidence units. Each generated QA pair contains a question, a reference answer, a set of supporting evidence units, and a composition category label.

For each sample, we first determine the composition requirement according to the structure of its annotated evidence. The question generation process then selects the relevant evidence units and creates a question whose solution requires jointly reasoning over these units. We describe the generation process for each category below.

Aggregation (AG). For aggregation questions, we identify evidence units that contain repeated instances, events, actions, or attributes distributed across different temporal locations. The generator combines these evidence units into a global query, such as counting occurrences, comparing frequencies, or summarizing a set of observed entities or events.

Specifically, given a collection of evidence units $\{e_1, e_2, ..., e_k\}$ with the same semantic target, we generate questions whose answers require accumulating information from all relevant units. Evidence from only one temporal location should provide incomplete information. The reference answer is generated by aggregating the attributes or events contained in the complete evidence set.

Temporal Composition (TC). For temporal composition questions, we utilize evidence annotations containing explicit temporal locations and event relationships. The generation procedure selects multiple events with meaningful temporal dependencies and creates questions requiring models to recover their ordering or temporal evolution.

Given evidence units with timestamps $\{(e_i, t_i)\}$, we generate questions involving relations such as before, after, first, last, or process development. The reference answer is derived from the temporal ordering of the annotated events rather than from the recognition of individual events alone.

Entity State Continuity (EC). For entity-state continuity questions, we exploit annotations describing entity identity and state transitions across time. The generator first identifies an entity that appears in multiple temporal regions and then constructs questions requiring tracking its identity, attributes, locations, or states.

For an entity trajectory $\{(e_i, s_i, t_i)\}$, questions are generated to query how the entity changes, what state it maintains, or how its properties evolve throughout the video. The supporting evidence therefore consists of multiple observations of the same entity, and answering requires associating these observations rather than treating them as independent detections.

Causal Reasoning & Summary (CS). For causal reasoning and summary questions, we use evidence units describing multiple events, outcomes, or narrative transitions. The generation process selects evidence groups that jointly describe a higher-level explanation or global video understanding.

Questions are generated to ask why an observable outcome occurs, how preceding events contribute to later states, or what major changes happen throughout the video. Since formal causal identifica-

tion is not available from observational videos, CS questions focus on video-grounded explanatory and narrative relations supported by visible events.

Unification and Normalization. After category-specific generation, all QA pairs are converted into a unified open-ended format. Existing multiple-choice questions are transformed by removing answer options and preserving only the underlying information need. Generated questions are further checked to ensure that the reference answer cannot be obtained from a single evidence unit and that all required information is supported by the annotated evidence.

The final QA set therefore provides explicit alignment between questions and distributed evidence structures, enabling controlled evaluation of different evidence composition requirements.

C.4 Quality Control

We apply multiple filtering stages to ensure that each benchmark instance genuinely evaluates distributed evidence composition.

Evidence Sufficiency. Each question must require at least two temporally separated evidence units. Samples answerable from a single frame or short local clip are removed.

Evidence Dependency. We verify that multiple evidence units contribute complementary information rather than simply co-occurring in the video. Questions requiring only object recognition or isolated event detection are excluded.

Temporal Distribution. Evidence regions are required to be distributed across the video timeline. This prevents the benchmark from degenerating into short-video question answering despite long video inputs.

Answer Grounding and Ambiguity Checking. Reference answers are checked to ensure they are fully supported by visible video evidence. Questions with ambiguous wording, incomplete evidence, or multiple plausible interpretations are removed.

Diversity and Balance. Finally, we analyze the distribution of video duration, domains, composition categories, and evidence demand. We maintain diversity across task types while ensuring that all categories contain sufficient samples for controlled evaluation.

After quality control, DECBench contains 926 videos and 1,123 open-ended questions covering four distributed evidence composition requirements.

D Additional Experiment Results

This section provides additional results for the diagnostic analyses and ablation studies discussed in the main paper. We first compare the evidence demand of DECBench with existing long-video benchmarks using Minimum Evidence Count (MEC), then report the category-level breakdown of the fixed visual-budget experiment, and finally provide detailed ablations of SEEK's structured evidence selection and query-adaptive evidence-acquisition profile.

D.1 MEC Comparison

Table 5 compares the average Minimum Evidence Count (MEC) across DECBench and three established long-video benchmarks. DECBench has the highest average MEC, at 5.264, compared with 3.049 for Video-MME, 3.918 for LongVideoBench, and 4.030 for LVBench. Because MEC measures the cardinality of a minimum sufficient evidence set, this result indicates that DECBench questions typically require a larger number of visual observations to be jointly available. This higher evidence demand complements the benchmark's focus on temporally distributed evidence composition and helps distinguish it from long-video evaluation settings in which fewer observations may suffice.

D.2 Fixed Visual-Budget Breakdown

Table 6 reports the category-level results of the fixed visual patch-budget experiment with Qwen3-VL-8B. We control the total image-patch budget used by the Qwen visual preprocessing pipeline

Table 5: Average Minimum Evidence Count (MEC).

Benchmark	MEC
Video-MME	3.049
LongVideoBench	3.918
LVBench	4.030
DECBench	**5.264**

while varying the number of sampled frames and the per-frame patch allocation. Specifically, we compare 64 × 256, 32 × 512, and 16 × 1024, where the first number denotes the number of frames and the second denotes the per-frame patch budget. Thus, all three configurations have the same aggregate patch budget of 16,384. Importantly, these patch counts refer to the preprocessing-level visual allocation and should not be interpreted as the exact number of visual tokens ultimately passed to the language model.

The 64×256 allocation achieves the highest overall accuracy (24.13%), outperforming 32×512 (23.42%) and 16×1024 (22.08%), which shows that broader temporal coverage is generally beneficial under a fixed budget. The category-level results, however, reveal a clear coverage–fidelity trade-off. AG and TC benefit most from the broadest coverage, whereas EC improves as more visual tokens are allocated per observation and reaches its best accuracy with 16×1024. CS is comparatively stable and attains its highest score with the same high-fidelity allocation. These results support using task-dependent visual-budget allocation rather than a single fixed coverage–fidelity configuration.

Table 6: Qwen3-VL-8B performance (%) under a fixed aggregate visual patch budget of 16,384. Each allocation is reported as number of frames × per-frame patch budget.

Allocation	AG	TC	EC	CS	Overall
64×256	19.64	26.67	15.64	34.43	24.13
32×512	19.27	24.64	16.73	32.97	23.42
16×1024	17.09	18.33	18.55	34.80	22.08

D.3 Ablation

Table 7 provides the full category-level ablations for SEEK on DECBench. For structured evidence selection, starting from the evidence-relevance-only baseline, adding temporal coverage improves overall accuracy from 25.11% to 26.27%. Incorporating relation consistency further raises it to 27.25%, and adding redundancy suppression yields the full SEEK result of 27.78%. The improvements are also consistent across all four composition categories, indicating that the components provide complementary benefits rather than gains confined to a single task family.

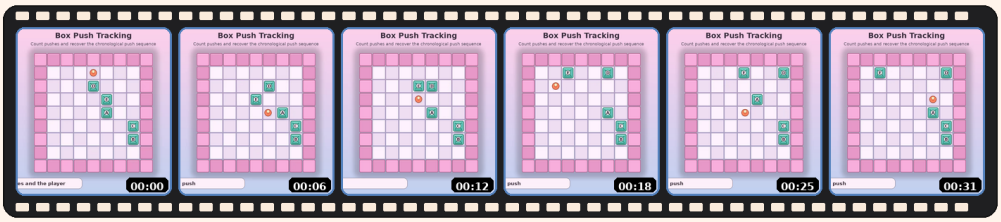
The evidence-acquisition-profile ablation further shows the importance of adapting the selection and allocation behavior to the query. Query-adaptive SEEK improves overall accuracy from 26.45% with a fixed profile to 27.78%, with gains in every category. The oracle-style GT task-specific profile reaches 28.14%, only 0.36 percentage points above Query-adaptive SEEK. Since the GT profile uses the ground-truth composition label whereas Query-adaptive SEEK infers its profile from the question alone, the small remaining gap suggests that query-conditioned planning recovers much of the benefit of task-specific evidence acquisition without requiring task labels or other oracle information at inference time.

Table 7: Ablation studies of structured evidence selection and the evidence-acquisition profile on DECBench using Qwen3-VL-8B-Instruct. Scores are accuracies (%). The GT task-specific profile uses the ground-truth composition label only as an oracle-style reference, whereas Query-adaptive SEEK predicts the profile from the question alone.

Setting	AG	TC	EC	CS	Overall
Structured Evidence Selection					
Evidence only	19.64	28.33	16.73	35.53	25.11
+ Temporal Coverage	21.09	29.67	17.45	36.63	26.27
+ Relation	21.45	31.33	18.18	37.73	27.25
Full SEEK (+ Redundancy)	21.82	31.67	18.91	38.46	27.78
Evidence-Acquisition Profile					
Fixed Profile	20.00	30.00	17.82	37.73	26.45
GT Task-specific Profile	22.18	32.00	19.27	38.83	28.14
Query-adaptive SEEK	21.82	31.67	18.91	38.46	27.78

E Data Examples

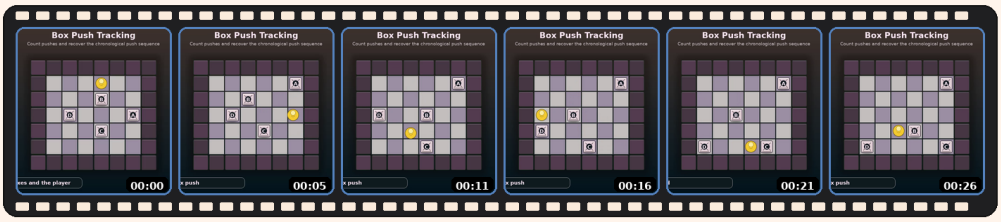
Aggregation



Q: The video shows a top-down box-pushing game with labeled boxes. The player walks around and sometimes pushes a box by one grid cell. Count how many actual box-push actions occur in the entire video. Return only a single integer.

A: 20

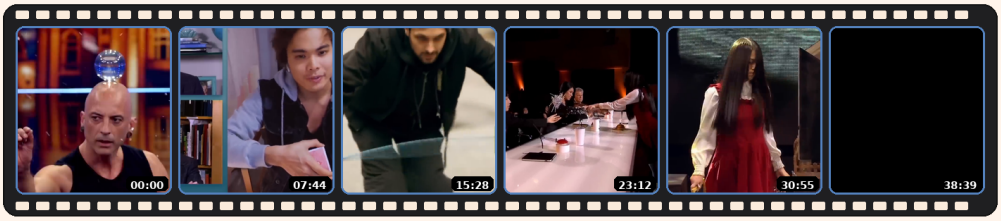
Aggregation



Q: The video shows a top-down box-pushing game with labeled boxes. The player walks around and sometimes pushes a box by one grid cell. Count how many actual box-push actions occur in the entire video. Return only a single integer.

A: 18

Aggregation



Q: How many tricks are revealed in this video?

A: 11.

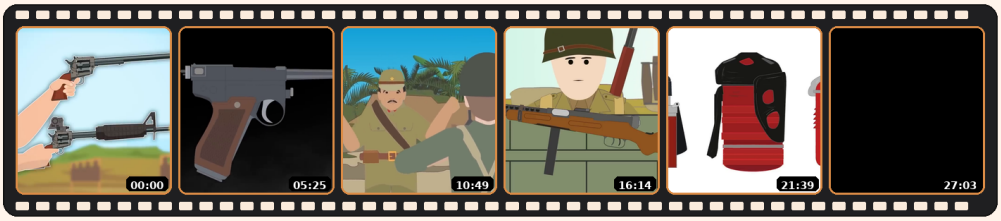
Aggregation



Q: In the video, where did the protagonist go to in Dubai during the VLOG shooting?

A: Dubai MALL, Dubai Old Town, Dubai Tower,

Temporal Composition



Q: Which of the following sequences of events is correct?
Events: a soldier wearing a steel helmet appears and is shooting towards the right side of the screen with his right hand holding the gun; a section of a gun appears with a hand holding a handgun and shooting at the section; a soldier wearing a green-yellow hat appears and is giving ammo to the caravan.

A: First, a soldier wearing a green-yellow hat appears and is giving ammo to the caravan, then a soldier wearing a steel helmet appears and is shooting towards the right side of the screen with his right hand holding the gun, and finally, a section of a gun appears with a hand holding a handgun and shooting at the section.

Temporal Composition



Q: What is the order in which the following is introduced in the video?
① How to supply oxygen
② The motion of ISS
③ Fireflies on ISS
④ The communication between ISS and The Earth

A: ①②③④.

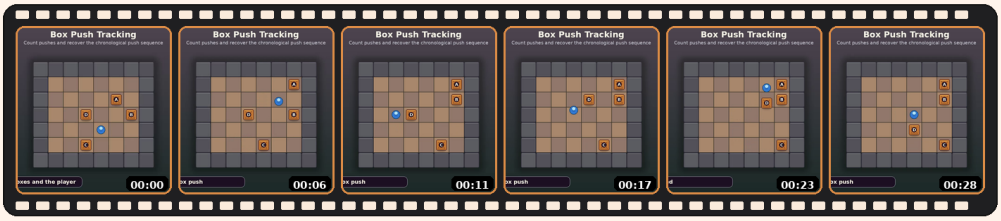
Temporal Composition



Q: In the video, when the protagonist is putting on makeup, in what order does she take the cosmetics?

A: Foundation, blush, eye shadow, eyeliner, mascara, lip gloss

Temporal Composition



Q: The video shows a top-down box-pushing game with labeled boxes. Below is a shuffled list of selected candidate push events from the video. Each candidate has the format EVENT_ID=BOX:DIRECTION. BOX is the visible box label. DIRECTION is one of U, D, L, R, meaning up, down, left, and right. Recover the chronological order of only these selected candidate events according to the video. Return only the EVENT_IDs in chronological order, separated by one space. Do not output box labels, directions, extra words, or punctuation. Shuffled candidate events: E04=A:R E02=C:L E01=B:U E03=A:U

A: E03 E04 E01 E02

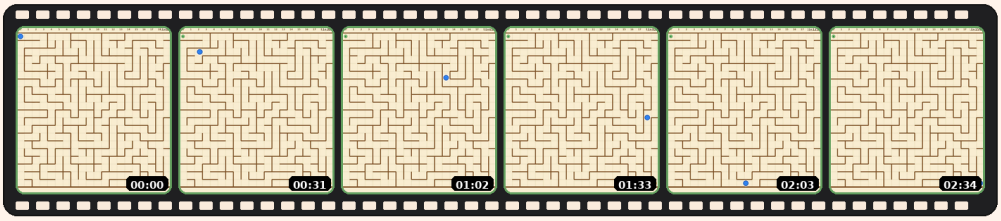
Entity State Continuity



Q: What notable similarities can be observed between the man and the woman's interactions with their cups throughout the video?

A: Both the man and the woman pick up their cups with their right hands, drink from them, and then place them back on the table.

Entity State Continuity



Q: The video shows a single highlighted moving ball traveling through a maze. Track this same ball during the interval from t=29.56s to t=36.13s. At the start of this interval, the ball is at cell (2,3); at the end of this interval, it reaches cell (5,0). Return only the ordered list of grid cells visited by the ball during this interval, including both the starting and ending cells. Use 0-indexed coordinates where x is the column index from left to right and y is the row index from top to bottom. Use exactly this format: [(x,y),(x,y),...].

A: [(2,3),(2,2),(3,2),(4,2),(5,2),(5,1),(5,0)]

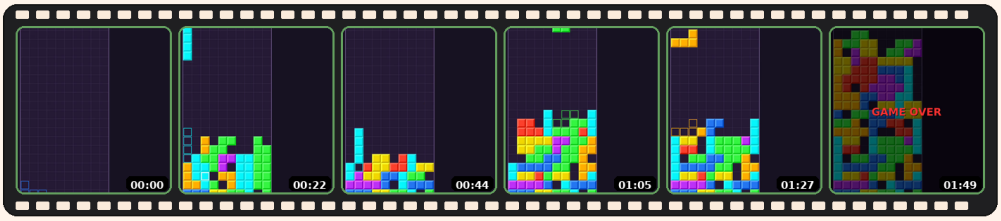
Entity State Continuity



Q: A person wearing a grass-woven headband is standing in the middle of the screen. On both sides, there are boxes filled with coins and jewels. In front of the boxes, there is a pile of money bags. When the man wearing the grass headband appears to the right of two long-haired dwarves, the background is red curtains over sunlit floor tiles. What change occurs to this man?

A: Black eyes turned into pink hearts with tears

Entity State Continuity



Q: A single square block is highlighted with a white outline after it lands. Track this exact block until it is cleared by a line-clear event. What is the block's board coordinate at the moment it is cleared? Use 0-indexed coordinates with row 0 at the top and column 0 at the left. Return only one coordinate in exactly this format: (row,col).

A: (17,2)

Causal Reasoning & Summary



Q: In the video, why does the protagonist, who is wearing a gray short-sleeved shirt and black trousers, put a brick on his belt?

A: In self-defense and revenge against anyone who splashed water on you.

Causal Reasoning & Summary



Q: Among all the actions in the video, which ones can be considered the most essential for achieving the desired outcome, and why?

A: The most essential actions for achieving the desired outcome are dipping the paint brush into the paint container and painting the edge of the window and the wall.

Causal Reasoning & Summary



Q: What explanation is given for the fact that the country mainly featured in the video continued to experience deflation in the early 2000s?

A: The deflationary mindset had become entrenched, leading to a self-fulfilling prophecy of low inflation expectations.

Causal Reasoning & Summary



Q: Describe the relationship between the measuring, writing, and cutting activities observed in the video, and explain their purpose in the context of the overall task being performed.

A: The measuring, writing, and cutting activities are all part of the process of measuring a wall and cutting a piece of wood to fit it. the measuring tape is used to measure the width of the wall, the paper is used to write down the measurements, and the wood cutter is used to cut the wood to the correct size.