SEAL: SEmantic Attention Learning for Long Video Representation

Lan Wang[1,2*] Yujia Chen[2] Du Tran[2] Vishnu Naresh Boddeti[1] Wen-Sheng Chu[2]
[1] Michigan State University [2] Google
{wanglan3, vishnu}@msu.edu {yujiachen, tranldu, wschu}@google.com

Abstract

Long video understanding presents challenges due to the inherent high computational complexity and redundant temporal information. An effective representation for long videos must efficiently process such redundancy while preserving essential contents for downstream tasks. This paper introduces **SE***mantic* **A***ttention* **L***earning (SEAL), a novel unified representation for long videos. To reduce computational complexity, long videos are decomposed into three distinct types of semantic entities: scenes, objects, and actions, allowing models to operate on a compact set of entities rather than a large number of frames or pixels. To further address redundancy, we propose an attention learning module that balances token relevance with diversity, formulated as a subset selection optimization problem. Our representation is versatile and applicable across various long video understanding tasks. Extensive experiments demonstrate that SEAL significantly outperforms state-of-the-art methods in video question answering and temporal grounding tasks across diverse benchmarks, including LVBench, MovieChat-1K, and Ego4D.*

1. Introduction

State-of-the-art video understanding models excell at short video tasks such as video classification [7, 13], temporal grounding [38] and action detection [10, 12], which involve videos lasting from *a few seconds* to *minutes*. However, their performance declines on hour-long videos [40, 41]. In contrast, a 10-year-old child can watch a full-length movie (*one to two hours*) and effortlessly answer questions at various levels of detail. This disparity between humans and machines emphasizes the foundamental challenges in long video understanding for AI models, including: (1) **Increased complexity**: Long videos require more computation and memory than current hardware can support for training or inference, (2) **Temporal redundancy**: Slow-changing scenes and objects introduce significant redundancy, and (3) **Cross-task generalization**: A robust representation must adapt to various tasks, from fine-grained fact retrieval to high-level reasoning. These challenges, which appear trivial for a child, remain challenging for AI models.

How does the human brain process long videos, particularly in addressing the above-mentioned challenges? First, rather than processing every pixel or frame, the brain selectively attends to new information to efficiently manage temporal redundancy [4, 22]. Second, humans process videos in an online fashion, continuously updating their understanding and memories as they watch, rather than deferring reasoning until the end. This continuous knowledge update, combined with selective attention, allows the brain to efficiently handle the complexity of long videos. Finally, attention dynamically shifts based on context. Without specific guidance, a child may focus on naturally engaging or memorable moments. However, when given specific questions in advance, attention becomes goal-oriented to seek relevant details while maintaining a broad understanding. This suggests effective representations should balance between task-specific focus and holistic understanding of the video to enable cross-task generalization.

Inspired by how humans process long videos, we introduce **SE**mantic **A**ttention **L**earning (SEAL), a novel unified representation designed to tackle the three key challenges in long video understanding. SEAL consists of two main steps: *Semantic Decomposition* and *Attention Learning*. In *Semantic Decomposition*, long videos are decomposed into three semantic entities such as scenes, objects, and actions, which are then treated as tokens. While the scene and object tokens represent static content assumed not to change rapidly, the action tokens are designed to capture the dynamic, fast-changing moments of the video. We note that these semantic tokens efficiently encode the essential information needed to answer "where", "what", or "how" questions about the videos. This decomposition significantly reduces complexity by allowing AI models to operate on a compact set of tokens instead of raw pixels or frames. Figure 1(a) illustrates a conventional uniformly sampled video $\mathcal{V}$, where redundant frames create cluttered visual information that hinders effective analysis for both models and humans. Figure 1(b) shows our semantic decomposition,

[*] This work was done as Lan Wang's internship project at Google.

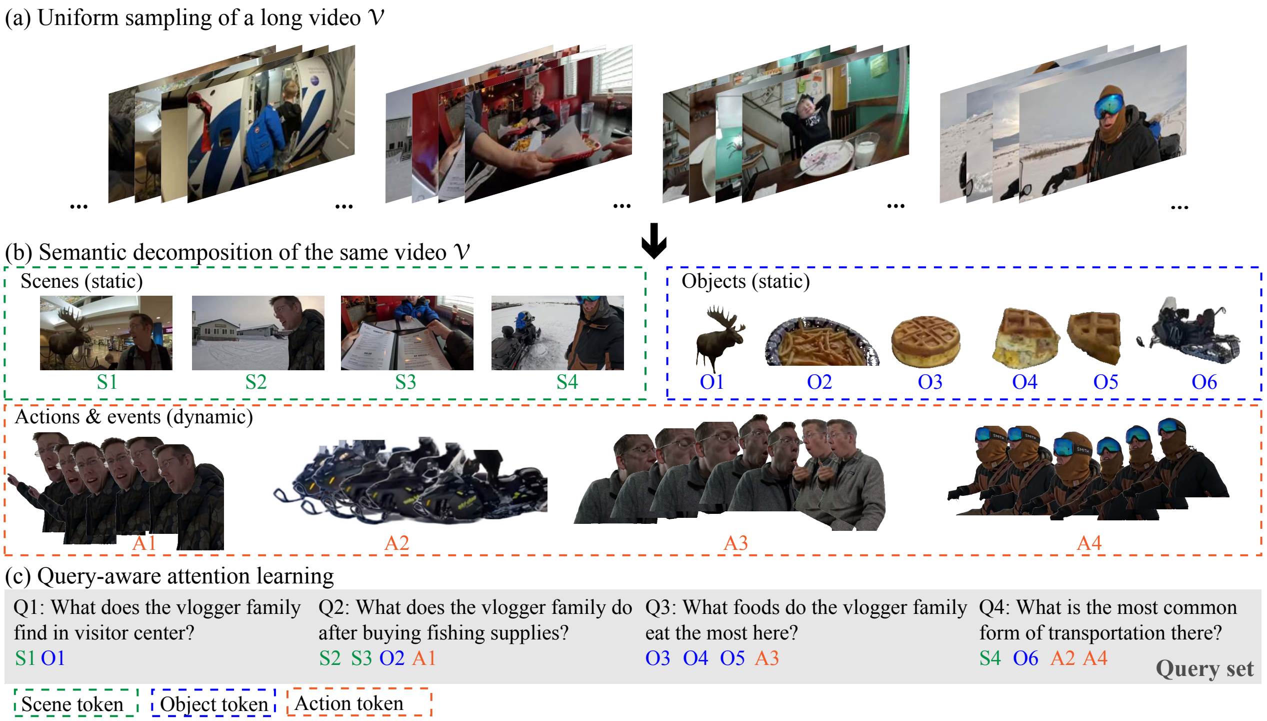

(a) Uniform sampling of a long video $\mathcal{V}$
...
(b) Semantic decomposition of the same video $\mathcal{V}$
Scenes (static)
S1
S2
S3
S4
Objects (static)
O1
O2
O3
O4
O5
O6
Actions & events (dynamic)
A1
A2
A3
A4
(c) Query-aware attention learning
Q1: What does the vlogger family find in visitor center?
S1 O1
Q2: What does the vlogger family do after buying fishing supplies?
S2 S3 O2 A1
Q3: What foods do the vlogger family eat the most here?
O3 O4 O5 A3
Q4: What is the most common form of transportation there?
S4 O6 A2 A4
Query set
Scene token
Object token
Action token

Figure 1. **Long Video Representation with Semantic Attention Learning (SEAL):** (a) Conventional uniform sampling results in redundant and cluttered visual information, making it difficult for both AI models and human brains to process efficiently. (b) Decomposing long videos into semantic entities such as scenes, objects, and actions reduces temporal redundancy, thus making model training and inference more efficient. In this example, the long video $\mathcal{V}$ is decomposed into 4 scene tokens (S1–S4), 6 object tokens (O1–O6), and 4 action/event tokens (A1–A4). (c) Query-aware attention learning module improves downstream task performance by focusing on relevant information rather than processing everything. Queries (Q1–Q4) are shown with their most relevant tokens. (best viewed in color)

breaking down the video into distinct scenes, static objects, and dynamic actions. In Attention Learning, we formulate a subset selection problem that maximizes the query relevance while ensuring token diversity. This step not only mitigates redundancy but also enhances cross-task generalization by prioritizing the most informative tokens. Figure 1(c) illustrates our attention learning module with four different queries and their selected tokens that capture relevance and diverse video content. Finally, SEAL is designed to work for both *global* and *streaming* modes, enabling it to process arbitrarily long videos. Extensive ablations and experiments confirm SEAL's superior performance over existing methods. Our key contributions include:

- We introduce SEAL, a novel unified representation for long videos by decomposing them into three semantic tokens, namely scenes, objects, and actions.
- Our attention learning module reduces temporal redundancy while supporting strong cross-task generalization. We show SEAL works in both global and streaming modes, making it adaptable to arbitrarily long videos.
- SEALoutperforms state-of-the-art methods on various long video understanding tasks and benchmarks including: video QA (MovieChat-1K [33], LVBench [40]), and egocentric video grounding (Ego4D [8]).

2. Related Work

Recent approaches have unified various video understanding tasks by framing them as video QA tasks [40], leveraging the capabilities of LLMs. However, fine-tuning with task-specific vision heads continues to offer advantages in memory efficiency and task-specific performance [29], particularly for temporal grounding. In this section, we review advancements in Video Question Answering (QA) and temporal grounding for long video understanding.

Video QA for long videos. The main challenge for long video QA is the memory constraint. He *et al*. [9] introduced a sequential framework that uses a memory bank to enhance long-term comprehension. Song *et al*. [33, 34] integrated video foundation models with LLMs through a memory mechanism inspired by the Atkinson-Shiffrin model, reducing computational complexity while preserving long-term memory. Another line of work improves efficiency by decomposing video content. Rui *et al*. [30] employed Memory-Propagated Streaming Encoding to segment videos into short clips, with Adaptive Memory Selection enhancing response accuracy by identifying question-relevant memories. Min *et al*. [24] introduced a multi-stage, training-free framework, emphasizing task decomposition into parsing, grounding, and reasoning stages. More recently, Weng *et al.* [41] proposed a hierarchical framework that encodes local features and integrates global semantics for detailed comprehension of extended video content. To further reduce model's hallucination in QA, Sun *et al.* [35] proposed a question-guided pipeline by focusing on relevant frames and controlled answer generation.

Temporal localization for long videos. Recent research in temporal localization for long videos has explored two primary directions: **LLM-based** and **non-LLM-based** approaches. With the advances in LLMs, researchers have expanded its use beyond traditional VQA, leveraging the capabilities for temporal grounding tasks, with a primary focus on enhancing localization accuracy. Ren *et al.* [32] proposed a timestamp-aware model that aligns visual content with temporal cues, enabling adaptive processing of sequential events for tasks like localization. Similarly, Fan *et al.* [5] introduced a memory-enhanced framework that captures contextual information across video segments, improving temporal and spatial reasoning. Korbar *et al.* [15] developed a text-guided resampling mechanism that dynamically selects video segments, focusing on relevant scenes to enhance temporal and spatial comprehension.

In contrast, non-LLM-based approaches typically rely on training a regression layer or decoder for temporal localization, with efforts concentrated on refining visual features and multimodal fusion to improve alignment. For instance, Hou *et al.* [11] introduced a hierarchical framework that combines coarse scanning with fine-grained alignment to optimize both precision and efficiency in localizing target moments. Pan *et al.* [28] applied a coarse-to-fine pipeline for single-pass temporal grounding that improves both efficiency and alignment. Additionally, Mu *et al.* [26] proposed a cost-effective late fusion approach paired with a video-centric sampling scheme to improve scalability.

Unlike prior works that focus on specific tasks, SEAL proposes a unified and generic framework for long video understanding, capable of adapting to various prediction heads and tasks. Additionally, SEAL supports both global and streaming modes, making it adaptable to arbitrarily long videos.

3. Semantic Decomposition and Attention Learning

Let $\mathcal{V} = \{v_i\}_{i=1}^{T_V}$ be an arbitrarily long untrimmed video, where $v_1, \ldots, v_{T_V}$ denote the sequence of T_V frames forming the video. Let q be a query from long video understanding tasks, comprising a sequence of l_q tokens. The query q may take different forms depending on the task, such as natural language text for video question and answering (*e.g.*, MovieChatQA [33], LVBench [40]) or visual/text template or action label for episodic memory tasks in egocentric video understanding (*e.g.*, Ego4D [8]). Our method establishes a unified video representation to generalize across these diverse long video understanding tasks. An overview of our approach is presented in Figure 2.

3.1. Semantic Decomposition of Long Videos

The main challenge in long video representation lies in capturing diverse content within limited memory. Conventional methods resort to frame sampling [42, 44] or maintaining a memory bank that merges similar frames [33, 34]. However, these approaches can vary greatly across tasks. We propose a novel decomposition approach that structures long videos into three distinct token types representing different type of semantic entities: (1) Scene tokens $\mathbf{T}_{\text{scene}}$ capture background context, providing essential cues about the environment. (2) Action tokens $\mathbf{T}_{\text{action}}$ represent moving elements, focusing on temporal information such as motions, activities, or events. (3) Object tokens $\mathbf{T}_{\text{object}}$, highlight key static elements relevant to specific tasks. With this structured tokenization, we create a unified, compact, task-agnostic representation that minimizes the need for redundant and dense video storage while preserving a comprehensive understanding of long video content.

Scene Tokens. Any frame in a video can serve as a scene token because it captures the environment where it was recorded, *e.g.*, indoor gym, outdoor mountain, etc. Although shot boundary detection could split a long video into shots with one scene token per shot, this approach has two drawbacks: (1) Shot detection algorithms are often imperfect, with no opportunity to correct once shot boundary is determined. (2) Shot detection often ignores the specific query and downstream tasks, and thus could cause sub-optimal performance. To overcome these drawbacks, we propose a two-step approach. First, we over-sample scene tokens, and then later perform an attention learning step to maximize query relevance and token diversity. Specifically, we uniformly sample N_{scene} frames to capture a diverse background. These pre-sampled scene tokens, denoted as $\mathbf{T}_{\text{scene}} = [t_i^{scene}]_{i=1}^{N_{\text{scene}}}$, will undergo another round of attention learning, detailed in Section 3.2.

Action Tokens. The purpose of the action token is to capture temporal information of moving objects such as low-level motions, activities, and events. We begin by using a class-agnostic object tracker, *e.g.*, SAM-2 [31], to extract multiple initial dynamic tracklets. Tracklets shorter than L_{min} are discarded, while those longer than L_{max} are split into multiple tracklets with length of L_{max}. For each tracklet, we take the spatial union of bounding boxes across frames, allowing the dynamic token to capture not only motion information but also the spatial movement of people and objects. This process yields N_{tracklet} tracklets, denoted as $\mathbf{T}_{\text{action}} = \{\tau_{\text{dynamic}}^i\}_{i=1}^{N_{\text{tracklet}}}$.

Object Tokens. For the object tokens, we utilize a class-agnostic grouping method such as SAM [14] to generate masks for all objects in each frame. This class-agnostic

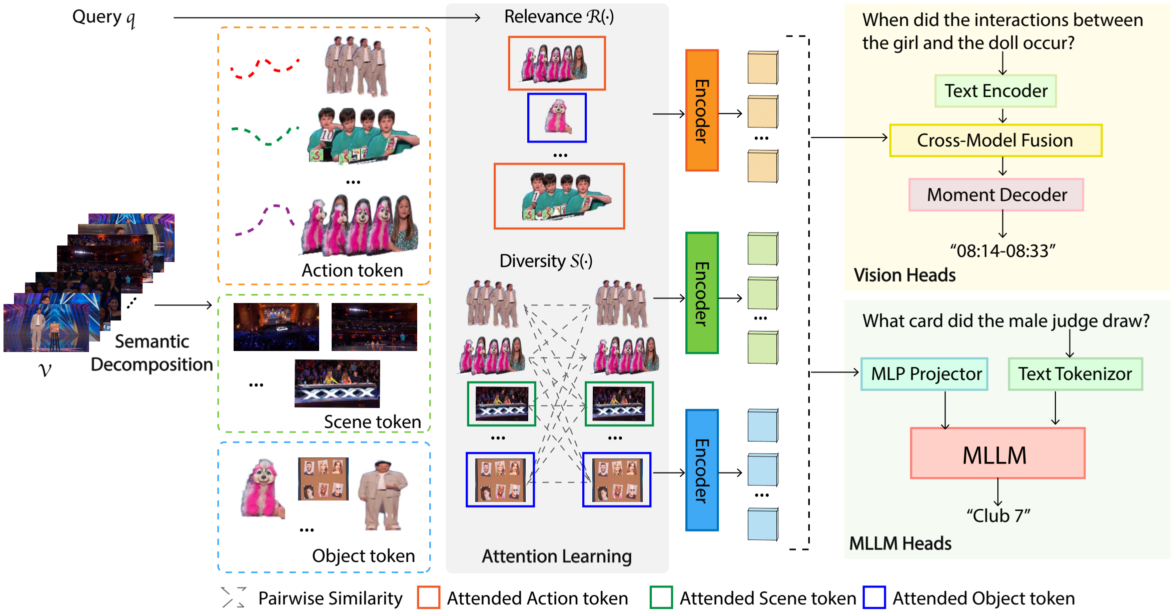


Figure 2. **SEAL Overview**. During *semantic decomposition*, a long video $\mathcal{V}$ is decomposed into semantic tokens representing scenes, objects, and actions. Then, during *attention learning*, these tokens and the query q, are optimized for query relevance $R(\cdot)$ and token diversity $S(\cdot)$. The resulting attended token subset is then passed to a vision or an MLLM head for predictions.

segmentation approach enables comprehensive object information capture. Specifically, we apply this process on N_{key} key frames. On each of the N_{key} frames, we apply SAM to obtain in total N_{object} object masks, which we denote as $\mathbf{T}_{\text{object}} = \{\mathbf{M}^i_{\text{object}}\}_{i=1}^{N_{\text{object}}}$. Key frame selection varies by task and is detailed in the following sections.

3.2. Attention Learning

Although the long video is decomposed into different token representations, the resulting tokens are still redundant. To address this, we propose a sampling approach that balances between query-relevance and token-diversity, denoted as Attention Learning. Specifically, we formulate our sampling as an optimization problem with two main objectives: query-relevance and token-diversity.

$$\begin{aligned} T_s^* &= \underset{T_s \subset T_G}{\arg\max}\, F_s(T_s | T_G, q) \\ &= \underset{T_s \subset T_G}{\arg\max}\, \alpha \sum_{t_s \in T_s} R(t_s, q) + (1-\alpha) \sum_{t_i, t_j \in T_s, i \neq j} \frac{1}{S(t_i, t_j)}. \end{aligned}$$

We aim to solve a subset selection problem: find a fixed-size subset $T_s \subset T_G$ that maximizes the objective function $F_s(T_s|T_G, q)$. Here, T_G represents the set of all tokens used, which can be any set of tokens such as scene $\mathbf{T}_{\text{scene}}$, object $\mathbf{T}_{\text{object}}$, action $\mathbf{T}_{\text{action}}$, or a combination of all ($\mathbf{T}_{\text{scene}} \cup \mathbf{T}_{\text{object}} \cup \mathbf{T}_{\text{action}}$). q denotes the query being asked by the downstream understanding task. And, T_s^* denotes the optimal subset that maximizes F_s. F_s is decomposed into two terms: $R(\cdot)$ measures the relevance between a visual token and the query which we compute by encoding them with the BLIP-2 model [17] and calculating their cosine similarity; and $S(\cdot)$ calculates the cosine similarity of the paired tokens. We note that the first term maximizes the relevance between selected tokens and the query while the second term enforces token diversity (via minimizing the token pairwise similarity). Finally, α is a hyper-parameter to balance between query-relevance and token-diversity.

3.3. Streaming and Global Mode

Our proposed method offers two ways of representing long videos: *streaming* and *global*. In the global mode, the model fully "watches" (or processes) the entire video and then provides a single representation. In contrast, in the streaming mode, the model processes the video buffer by buffer and provides an updated representation at any given time step. The partly-observed video representation can change over time as the video progresses. This setting simulates the situation when you watch a movie together with a child and have interactions with her or him as the movie is still going on.

In the global model, all tokens undergo a single optimization through Attention Leaning, resulting in T_{sub}, which includes k tokens. This T_{sub} is regarded as the representation for the entire video. We note that the global mode may not scale well with arbitrary long videos because all tokens cannot be fit into a limited memory for sampling. One can opt to use more aggressive uniform temporal presampling to reduce the numbers of tokens before Attention Learning. This workaround can bypass the memory limitation, but may also lead to sub-optimal solutions due to missing important tokens due to uniform sampling.

Alternatively, we propose an online streaming approach for representing partly-observed videos. Specifically, we use a fixed-size sliding window with a size set to the maximum number of tokens l allowed by memory capacity, denoted as T. At each step t, we apply Attention Learning to the union set of the tokens in the current window T_t and the previous selected subset of tokens T_{sub}^{t-1} and obtain the representation of the video at time T_{sub}^{t}. At the beginning, the selected subset is set to empty ($T_{\text{sub}}^{0} = \emptyset$).

$$T_{\text{sub}}^{t} = \text{Attention\_Learning}(T_t \cup T_{\text{sub}}^{t-1})\ \forall t > 0. \quad (1)$$

This streaming mode allows us to use T_{sub}^{t} as the partly-observed representation of the video and can be fed into any prediction head for video understanding tasks. As an immediate benefit of the streaming mode, our proposed representation now can handle arbitrary long videos.

3.4. Prediction Heads

Our unified representation is adaptable to most long video understanding tasks using different prediction heads. In this paper, we demonstrate two specific use cases of our representation: one is used with the traditional vision head for video temporal grounding and the other one is with the Multimodal LLMs (MLLM) head for video QA.

Temporal Grounding with Vision Heads. Given a query q, the task is to locate the start and end times, t_{start} and t_{end}, where the answers could be deduced. We first encode the sampled tokens T_{sub} and the query using encoders $\mathbb{E}_V$ and $\mathbb{E}_q$ to obtain embeddings for each video token $z_v^i = \mathbb{E}_V(T_{\text{sub}^i})$ and query $z_q = \mathbb{E}_q(q)$. The cross-modal fusion is then performed to obtain the fused representation:

$$z_{\text{joint}}^{i} = \text{CrossModalFusion}(z_v^i, z_q) \quad (2)$$

Finally, a moment decoder is applied to predict the start and end time for the query q. The moment decoder includes a classification head and a regression head. The classification head is used to predict the score p_{score}^{i} of each token, while the regression head predicts the normalized distances $(d_{\text{start}}^{i}, d_{\text{end}}^{i})$ from each token to the moment boundaries:

$$p_{score}^{i}, (d_{\text{start}}^{i}, d_{\text{end}}^{i}) = \text{MomentDecoder}(z_{\text{joint}}^{i}). \quad (3)$$

The regression and classification heads are optimized using an IoU distance and a focal loss as used in [25]. The final moment is calculated as $t_{start}^{i}, t_{end}^{i} = (t_i - d_{\text{start}}^{i}) \times L_V, (t_i + d_{\text{end}}^{i}) \times L_V$, where L_V is the length of the input video. The proposed unified representation enables efficient handling of long video sequences and allows for localization related tasks.

Video QA using MLLM Heads. The proposed representation can also be connected to an MLLM head, making it applicable to various video QA-related tasks, such as reasoning, understanding, and summarization. Moreover, some grounding tasks can also be addressed in a QA format. Specifically, z_v is projected through an MLP to a visual token T_v^{MLLM} that the MLLM can interpret, which is then input into the MLLM along with text tokens T_{text}. Based on different benchmarks, the MLLM performs multiple-choice or open-ended answering.

4. Experiments

4.1. Implementation Details

Datasets and metrics. We evaluate SEAL on three datasets, each selected for its relevance to long video understanding on different capabilities. *LVBench* [40] contains 1,549 QA pairs across six tasks, with videos averaging 4,101 seconds (approximate 1 hour 8 minutes). Each question presents a single-choice format with four options. Accuracy serves as the evaluation metric for individual tasks as well as overall performance across all tasks. We primarily focus on this dataset due to its emphasis on hour-long videos. *Moviechat-1K* [33] includes 1,000 video clips with dense captions spanning 15 categories, averaging 564 seconds (about 10 minutes). The benchmark employs LLMs, specifically GPT-3.5 [27], to evaluate the quality of generated answers. A rating ranging from 0 to 5 is used to compute the overall score, while a binary preference from the LLM is used to calculate the accuracy. *Ego4D-NLQ* [8] is a part of the Ego4D Episodic Memory challenge for Natural Language Queries (NLQ) task. This dataset requires localizing a temporal window where the answers can be deduced from untrimmed egocentric videos. It contains 1,259 videos, averaging 10 minutes each. Our experiments apply memory constraints to simulate long video scenarios. Metrics include Top-1 and Top-5 recall at various thresholds.

Experiment setup. We fine-tune the projection layers and Q-former [17] to adapt to different types of tokens for 20 epochs using the training split of MovieChat-1K. For Ego4D-NLQ, we use the training split to finetune the vision heads for 7 epochs. The AdamW optimizer [21] is employed with default beta values of (0.9, 0.999) and a weight decay of 0.05. For token extraction, we use SAM2 [31] to obtain object tokens, YOLOv10-X [37] with BoT-SORT [1] for action tokens. Scene, action, and object tokens are extracted at 8, 10, 1 FPS for MovieChat-1K, Ego4D-NLQ and LVbench. For the LVBench, we follow the settings of InternVL2 and utilize Yi-34B [43]. For MovieChat-1K, we adopt the same settings as [33] and use Vicuna-7B as [33]. For Ego4D-NLQ, we adhere to the settings specified in SnAG [25] and use EgoVLP [19] as vision encoder. We use 0.9 as the default value for the only hyper-parameter α.

Model	LLM Size	Overall (%)	KIR (%)	EU (%)	Sum (%)	ER (%)	Rea (%)	TG (%)
Qwen2-VL [39]	72B	41.3	38.3	41.1	**46.6**	38.0	**46.5**	**41.4**
InternVL2 [3]	34B	39.6	43.4	39.7	41.4	37.4	42.5	31.4
LLaVA-NeXT [45]	34B	32.2	34.1	31.2	27.6	30.1	35.0	31.4
Oryx [20]	34B	30.4	32.1	29.2	27.6	30.1	34.0	29.1
PLLaVA [42]	34B	26.1	26.2	24.9	25.9	25.0	30.0	21.4
SEAL (Ours)	34B	**45.9**	**51.5**	**41.3**	39.7	**47.9**	43.3	32.3

Table 1. **Comparison with state-of-the-art on LVBench**. Our method achieves the highest overall score, with notable gains in Key Information Retrieval (KIR) and Entity Recognition (ER), demonstrating the effectiveness of our representation in locating key information and entities by eliminating redundancy. The best methods are highlighted in **bold**, and the second-best are underlined.

		R@1			R@5		
Model	#Token	0.3	0.5	Avg	0.3	0.5	Avg
SnAG	all	15.72	10.78	13.25	38.39	27.44	32.92
SnAG	450	13.44	9.23	11.34	34.02	23.04	28.53
SEAL	450	**13.78**	**9.26**	**11.52**	**34.79**	**23.10**	**28.95**
SnAG	200	10.03	6.35	8.19	26.56	16.90	21.73
SEAL	200	**10.83**	**7.06**	**8.95**	**27.39**	**17.41**	**22.40**

Table 2. **Comparisons with SoTA methods on Ego4D-NLQ**. Quantitative results for temporal grounding on Ego4D episodic memory Natural Language Queries (NLQ) task show our method, SEAL, consistently outperforms SnAG [25] in all metrics under memory constraints with varying number of tokens.

	Methods	Accuracy(%)	Score
Zero-shot	VideoChat [18]	61.0	3.34
	VideoLLaMA [44]	51.4	3.10
	Video-ChatGPT [23]	44.2	2.71
	MovieChat [33]	67.8	3.81
Supervised	TimeChat-Hal [35]	73.8	3.58
	HERMES [6]	84.9	**4.40**
	SEAL (Ours)	**86.8**	4.35

Table 3. **Comparison with SoTA methods on MovieChat-1K**. SEAL outperforms the second best method, HERMES, by 1.9% on accuracy while being comparable on the score metric.

4.2. Comparison with State-of-the-arts

We conduct quantitative analysis on LVBench dataset which contains videos on lengths averaged more than 1 hour. Additionally, we demonstrate the generalizability of the proposed unified representation using MovieChat-1K, which has open-ended QA questions on a variety of scenarios, and Ego4D-NLQ, where we use traditional vision decoders (LLM-independent) for the temporal grounding task.

LVBench. Table 1 shows the accuracy of our method on different categories of LVBench dataset [40], demonstrating that SEAL achieves the highest overall score, notably outperforming even larger models like Qwen2-VL-72B [39] by 4.6%. Our method excels in Key Information Retrieval (KIR) and Entity Recognition (ER), outperforming the strongest alternatives by 8.1% and 5.1% respectively. These results highlight that our action and object tokens effectively locate key information and entities by removing redundancies. While larger LLMs often exhibit better performance, this analysis suggests that it is not the only determinant, and a unified representation like ours can achieve state-of-the-art results with fewer parameters.

Figure 3 presents a qualitative analysis of SEAL on LVBench, demonstrating that our approach is able to pay attention to relevant semantic tokens and make correct answers to different types of questions. SEAL accurately locates relevant tokens, *e.g.*, identifying the royal family's stool color (Q1.a), counting meals eaten (Q2.a), or determining scene or location (Q2.b). We also show a failure case (Q2.c), where the nuanced "why" question requires complicated causal relations of different scenes.

Moviechat-1K. Table 3 compares our method with state-of-the-art methods on MovieChat-1K. We focus on evaluating performance in the global mode, which assesses the model's ability to comprehend information from the entire video, rather than the breakpoint mode, which primarily evaluates its ability to answer questions related to specific timestamps. Our method surpasses all the alternatives, achieving the highest accuracy and a strong score. Notably, unlike LVBench, which employs multi-choice questions, this dataset evaluates generated answers using an off-the-shelf LLM. This highlights the ability of our learned representations to effectively transfer to downstream generative tasks, enabling generating accurate and detailed responses.

Ego4D-NLQ. We also evaluate our approach on the Ego4D-NLQ task, which differs from previous datasets as it uses a decoder to locate events for temporal grounding, rather than relying on an LLM to generate text-based answers. Table 2 shows the results under memory-constrained conditions, where we limit token numbers to simulate real-world memory limitations in long video understanding. Our method consistently achieves the highest recalls across various thresholds. With further memory constraints (*e.g.*, reducing tokens to 200), SEAL outperforms the current SoTA method, SnAG [25], by a even larger margin. This demonstrates the robustness of our unified representation on various downstream applications for long video understanding.

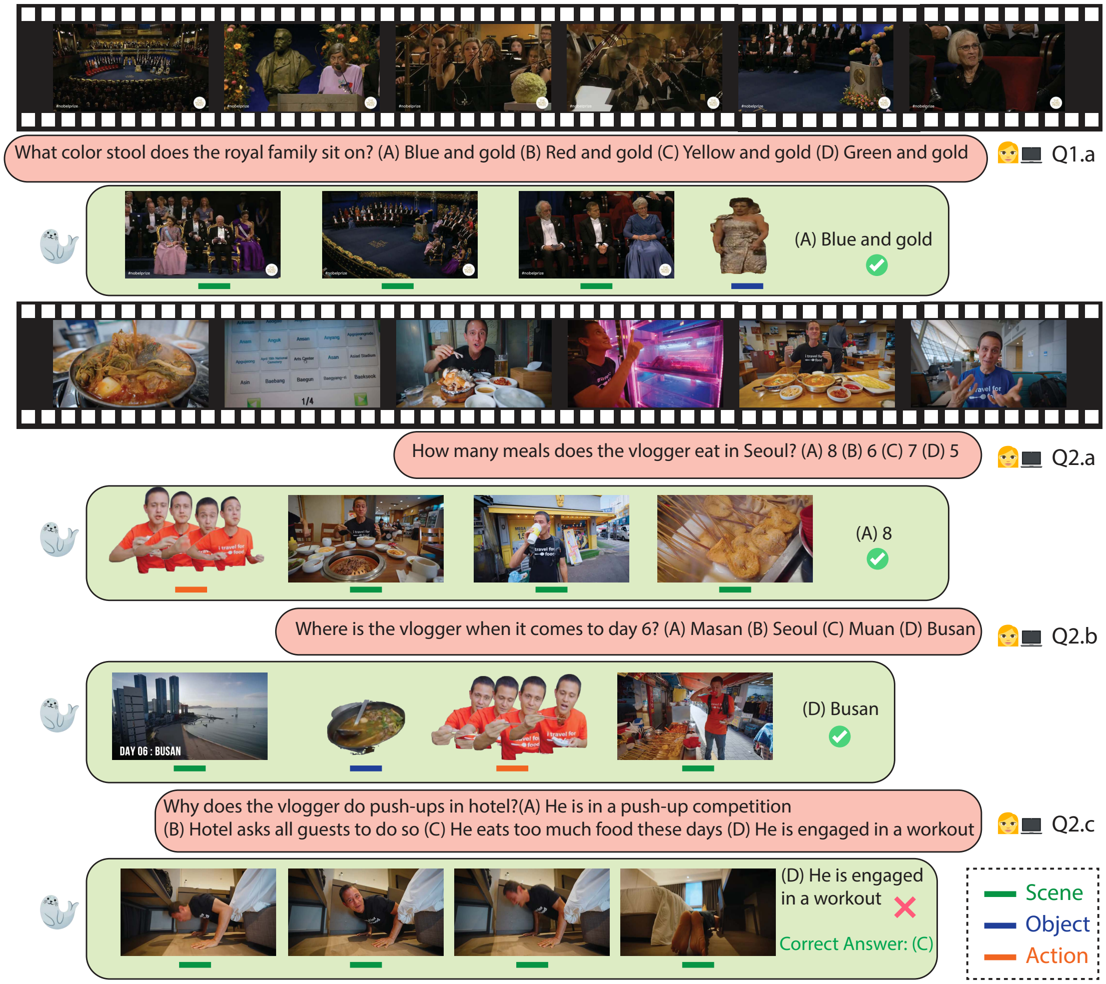

What color stool does the royal family sit on? (A) Blue and gold (B) Red and gold (C) Yellow and gold (D) Green and gold
Q1.a
(A) Blue and gold ✓
How many meals does the vlogger eat in Seoul? (A) 8 (B) 6 (C) 7 (D) 5
Q2.a
(A) 8 ✓
Where is the vlogger when it comes to day 6? (A) Masan (B) Seoul (C) Muan (D) Busan
Q2.b
DAY 06 : BUSAN
(D) Busan ✓
Why does the vlogger do push-ups in hotel?(A) He is in a push-up competition
(B) Hotel asks all guests to do so (C) He eats too much food these days (D) He is engaged in a workout
Q2.c
(D) He is engaged in a workout ✗
Correct Answer: (C)
Scene
Object
Action

Figure 3. **Qualitative results on LVBench**. Two long videos visualized with questions, multiple choice options, and SEAL predicted answers. SEAL attends to relevant entities such as “royal family” and “stool” (Q1.a), different “meals” and “drinks” (Q2.a), “scene” and “location” (Q2.b) and correctly answers these questions. Although attending to relevant “push-up” activity (Q2.c), SEAL fails to predict the right answer due to the challenging in the causal reasoning question.

4.3. Ablation Study

Semantic decomposition. We analyze the impact of the three proposed token types on LVBench (Table 4). To prevent LLMs from leveraging prior knowledge to “guess” answers using only the text query, we present the baseline that uses random token inputs, isolating the LLM’s performance. Results indicate that using each token type individually significantly improves performance, highlighting the value of each token’s contribution. Notably, scene tokens provide the most substantial improvement, aligning with prior research [2, 36, 41] that emphasizes frame-based scene tokens. Furthermore, different token combinations yield additional gains: adding action tokens enhances Key Information Retrieval (KIR), Reasoning (Rea) and Temporal Grounding (TG) tasks by capturing temporal dynamics in the queries, while object tokens boost Entity Recognition (ER) performance by retaining detailed object-specific information. We observe high variance in the Summarization (Sum) accuracy due to the small number of questions in this category and the randomness of LLM. Similar improvements have been observed on the MovieChat-1K in Table 5, where the accuracy improved by 10.45% with action and object tokens. Ultimately, using all three token types together achieves the highest performance, demonstrating the complementary strengths of scene, action, and object tokens as a unified representation in long video understanding.

We observe that the latency bottleneck mostly comes from the tracker when processing multiple objects. Table 6 presents an analysis of different trackers for extracting action tokens. While YOLO+BoT-SORT operates in a much higher frame rate (14 FPS), it lacks open-set detection capabilities. On the other hand, SAM2 can detect all objects within a scene but operates significantly slower (8FPS). Our findings indicate that the proposed action tokens does not

Model	Overall	KIR	EU	Sum	ER	Rea	TG
Random tokens	24.8	24.7	25.2	34.5	23.5	26.4	20.9
Action only	34.0	33.3	31.7	36.2	34.3	35.8	33.6
Object only	33.9	34.4	32.6	31.0	33.4	40.8	29.5
Scene only	42.8	50.5	40.6	25.9	44.3	41.3	27.2
Scene+Object	43.4	47.4	40.0	24.1	47.6	42.3	35.5
Scene+Action	44.4	**53.3**	40.0	24.1	46.2	**43.3**	**38.2**
SEAL (Ours)	**45.9**	51.5	**41.3**	**39.7**	**47.9**	**43.3**	32.3

Table 4. **The effects of different types of tokens on LVbench**. Using any type of tokens outperforms the random baseline, while applying all three types of tokens brings the best performance.

Method	Scene	Scene+Object	Scene+Action	SEAL
Accuracy	76.33	79.49	81.46	**86.78**
Score	4.15	4.26	4.28	**4.35**

Table 5. **The effects of different types of tokens on MovieChat-1K**. Results are on the test set with our global inference mode. Using all three types of tokens provides the best accuracy.

require capturing all the objects in the scene. In practice, YOLO-X pretrained on COCO, with BoT-SORT, effectively captures essential information for action token extraction, maintaining a balance between performance and efficiency.

Attention learning. Table 6 highlights the effectiveness of query relevance and token diversity terms in Section 3.2. When $\alpha = 0$, the optimization focuses on the diversity term to capture extensive contextual information for long-video tasks such as Summarization (Sum) and Reasoning (Rea), achieving reasonable results comparable to other state-of-the-art approaches in Table 1. With $\alpha = 1$, the optimization prioritizes the relevance term to reduce redundancy by sampling tokens that closely align with the query, so performance can be improved for tasks like Key Information Retrieval (KIR). However, this approach loses global context and results in weaker summarization performance. Balancing both terms yields the best overall results.

Streaming mode. We compare the global mode and streaming mode of our method on LVBench. As discussed in Section §3, the streaming mode better simulates real-world scenarios by sequentially processing partial observations to aggregate information over arbitrarily long videos. Table 6 demonstrates that streaming mode performs worse on globally-dependent tasks such as Event Understanding (EU) and Key Information Retrieval (KIR). However, it excels in temporally-intensive tasks like Temporal Grounding (TG) and Reasoning (Rea). Notably, streaming mode still surpasses the most competitive baseline, even when the baseline uses a significantly larger LLM.

Computational Complexity. We provide an accuracy-efficiency trade-off comparison in Fig 4 where both SEAL and InternVL2 are with a varying number of tokens. FPS is calculated during inference that includes all the processing time from the raw video frames. SEAL reduces the reliance on densely sampled tokens and uses subject-level and motion aware representations/tokens to improve efficiency while maintain good accuracy. From Fig 4, at 10 FPS, SEAL is about 5% more accurate than InternVL2. When comparing at the same accuracy of 41.5% and 40%, SEAL is about 3x and 2x faster than InternVL2, respectively.

Model	Overall	KIR	EU	Sum	ER	Rea	TG
SEAL (Ours)	45.9	51.5	41.3	39.7	47.9	43.3	32.3
rep Yolo w/ SAM2	43.3	51.9	39.9	29.3	46.4	40.3	32.7
⇒ Streaming	44.2	50.9	39.7	37.9	45.0	44.3	34.1
Diversity only	38.7	40.4	38.8	40.0	38.7	40.9	33.7
Relevance only	42.0	48.5	39.9	27.6	44.0	37.3	31.8

Table 6. **Ablation studies on LVBench.** SEAL defaults to global mode with YOLO for efficiency, while SAM2 performs similarly. Global mode excels in long-context tasks, and streaming mode in temporally-intensive tasks. Combining Relevance and Diversity in attention learning achieves the best performance.

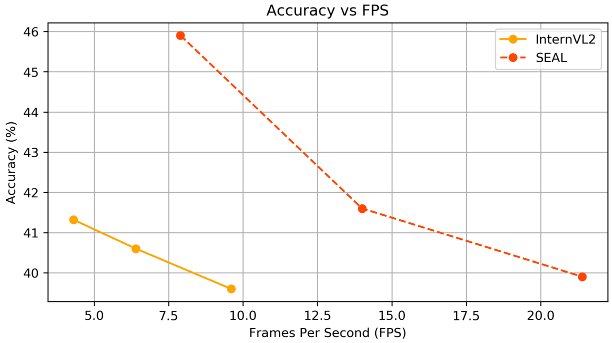


Figure 4. **Accuracy vs. efficiency trade-off on LVBench**. SEAL runs 2-3x faster than InternVL2 at the same accuracy, and is more accurate when compared at the same FPS.

5. Conclusion and Future Work

We propose SEAL, a novel unified representation for long video understanding that addresses computational complexity, temporal redundancy, and cross-task generalization. SEAL leverages semantic decomposition to break videos into scene, object, and action tokens, reducing redundancy and enabling efficient processing. It incorporates attention learning to balance query relevance and token diversity, enhancing performance across diverse tasks. SEAL achieves state-of-the-art results on benchmarks like MovieChat-1K, LVBench, and Ego4D-NLQ, demonstrating its versatility and effectiveness for long video understanding.

Limitation. The Attention Learning module is bounded by the memory constraint for the QP solver, making it not fully end-to-end trainable. For future work, we plan to integrate Attention Learning in our streaming mode for full end-to-end learning. Developing special prediction heads to solve causal reasoning [16], where MLLM heads showed limitations, is also an interesting area for future exploration.

Acknowledgements. This work was partially supported by the Office of Naval Research (award #N00014-23-1-2417). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of ONR.

References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Botsort: Robust associations multi-pedestrian tracking. *ArXiv*, abs/2206.14651, 2022. 5

[2] Minghao Chen, Fangyun Wei, Chong Li, and Deng Cai. Frame-wise action representations for long videos via sequence contrastive learning. In *CVPR*, 2022. 7

[3] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 6

[4] Meghomala Das, David M Bennett, and Gordon N Dutton. Visual attention as an important visual function: an outline of manifestations, diagnosis and management of impaired visual attention. *British Journal of Ophthalmology*, 91(11): 1556–1560, 2007. 1

[5] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *ECCV*, pages 75–92. Springer, 2024. 3

[6] Gueter Josmy Faure, Jia-Fong Yeh, Min-Hung Chen, Hung-Ting Su, Shang-Hong Lai, and Winston H. Hsu. Hermes: temporal-coherent long-form understanding with episodes and semantics. *arXiv preprint arXiv:2408.17443*, 2024. 6

[7] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 1

[8] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolár, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard A. Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the world in 3, 000 hours of egocentric video. In *CVPR*, 2022. 2, 3, 5

[9] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *CVPR*, 2024. 2

[10] Fabian Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 1

[11] Zhijian Hou, Wanjun Zhong, Lei Ji, Difei Gao, Kun Yan, Wing-Kwong Chan, Chong-Wah Ngo, Zheng Shou, and Nan Duan. Cone: An efficient coarse-to-fine alignment framework for long video temporal grounding. *arXiv preprint arXiv:2209.10918*, 2022. 3

[12] Haroon Idrees, Amir R. Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The THUMOS challenge on action recognition for videos "in the wild". *CoRR*, abs/1604.06182, 2016. 1

[13] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 1

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *ICCV*, 2023. 3

[15] Bruno Korbar, Yongqin Xian, Alessio Tonioni, Andrew Zisserman, and Federico Tombari. Text-conditioned resampler for long form video understanding. In *ECCV*, pages 271–288. Springer, 2024. 3

[16] Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *CVPR*, 2022. 8

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*. PMLR, 2023. 4, 5

[18] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wen Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 6

[19] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *arXiv preprint arXiv:2206.01670*, 2022. 5

[20] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024. 6

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[22] Steven J. Luck and Michelle Ford. On the role of selective attention in visual perception. *Proceedings of the National Academy of Sciences of the United States of America*, 95(3): 825–830, 1998. 1

[23] Muhammad Maaz, Hanoona Abdul Rasheed, Salman H. Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *AMACL*, 2023. 6

[24] Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In *CVPR*, pages 13235–13245, 2024. 2

[25] Fangzhou Mu, Sicheng Mo, and Yin Li. SnAG: Scalable and accurate video grounding. In *CVPR*, 2024. 5, 6

[26] Fangzhou Mu, Sicheng Mo, and Yin Li. Snag: Scalable and accurate video grounding. In *CVPR*, pages 18930–18940, 2024. 3

[27] OpenAI. Gpt-3.5, 2023. 5

[28] Yulin Pan, Xiangteng He, Biao Gong, Yiliang Lv, Yujun Shen, Yuxin Peng, and Deli Zhao. Scanning only once: An end-to-end framework for fast temporal grounding in long videos. In *ICCV*, pages 13767–13777, 2023. 3

[29] Baoqi Pei, Guo Chen, Jilan Xu, Yuping He, Yicheng Liu, Kanghua Pan, Yifei Huang, Yali Wang, Tong Lu, Limin Wang, et al. Egovideo: Exploring egocentric foundation model and downstream adaptation. *arXiv preprint arXiv:2406.18070*, 2024. 2

[30] Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. *arXiv preprint arXiv:2405.16009*, 2024. 2

[31] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos, 2024. 3, 5

[32] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *CVPR*, pages 14313–14323, 2024. 3

[33] E. Song, W. Chai, G. Wang, Y. Zhang, H. Zhou, F. Wu, X. Guo, T. Ye, Y. Lu, JN. Hwang, and G. Wang. MovieChat: From dense token to sparse memory for long video understanding. In *CVPR*, 2024. 2, 3, 5, 6

[34] Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. Moviechat+: Question-aware sparse memory for long video question answering. *arXiv preprint arXiv:2404.17176*, 2024. 2, 3

[35] Yiwei Sun, Zhihang Liu, Chuanbin Liu, Bowei Pu, Zhihan Zhang, and Hongtao Xie. Hallucination mitigation prompts long-term video understanding. *arXiv preprint arXiv:2406.11333*, 2024. 3, 6

[36] Reuben Tan, Ximeng Sun, Ping Hu, Jui-hsien Wang, Hanieh Deilamsalehy, Bryan A Plummer, Bryan Russell, and Kate Saenko. Koala: Key frame-conditioned long video-llm. In *CVPR*, 2024. 7

[37] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yolov10: Real-time end-to-end object detection. *ArXiv*, abs/2405.14458, 2024. 5

[38] Lan Wang, Gaurav Mittal, Sandra Sajeev, Ye Yu, Matthew Hall, Vishnu Naresh Boddeti, and Mei Chen. Protege: Untrimmed pretraining for video temporal grounding by video temporal grounding. In *CVPR*, 2023. 1

[39] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6

[40] W. Wang, Z. He, W. Hong, Y. Cheng, X. Zhang, J. Qi, S. Huang, B. Xu, Y. Dong, M. Ding, and J. Tang. LVBench: An extreme long video understanding benchmark, 2024. 1, 2, 3, 5, 6

[41] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *ECCV*, pages 453–470. Springer, 2024. 1, 3, 7

[42] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024. 3, 6

[43] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai. *ArXiv*, abs/2403.04652, 2024. 5

[44] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *EMNLP*, 2023. 3, 6

[45] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 6

SEAL: SEmantic Attention Learning for Long Video Representation

Supplementary Material

A. Additional Ablations

Streaming Window Size. Table 1 demonstrates the impact of different window sizes on performance in streaming mode. This ablation experiment simulates the behavior of streaming mode under varying memory constraints. The results show that the performance of streaming mode is optimal when the window size is set to 1000, demonstrating its ability to effectively balance memory usage and accuracy under this configuration.

Model	Overall	KIR	EU	Sum	ER	Rea	TG
Ours - 500	42.7	50.5	38.9	39.7	43.4	41.8	32.7
Ours - 1000	44.2	50.9	39.7	37.9	45.0	44.3	34.1
Ours - 2000	42.7	52.9	40.3	20.7	43.7	37.8	29.5

Table 1. **Ablation studies of Streaming Window Size on LVBench.** Streaming mode performs best when the window size is set to 1000.

Partially-Observed Videos. Table 2 presents the performance of different methods when only a portion of video is accessible including scenarios where only the first half or quarter of the video is available. This experiment simulates streaming mode, where the model receives only a portion of the video as input, evaluating its ability to answer questions under such constraints. The results show that our method significantly outperforms the uniform sampling approach of the InterVL2-40B model, highlighting the effectiveness of our relevant and diverse tokens.

Model	Observed	Overall	KIR	EU	Sum	ER	Rea	TG
InterVL2-40B	1	39.6	43.4	39.7	41.4	37.4	42.5	31.4
SEAL (Ours)	1	45.9	51.5	41.3	39.7	47.9	43.3	32.3
InterVL2-40B	1/2	35.7	34.4	34.2	37.9	35.7	40.3	28.2
SEAL (Ours)	1/2	41.6	50.9	37.9	41.4	41.9	39.8	29.5
InterVL2-40B	1/4	35.6	36.4	33.8	34.4	34.1	37.5	27.3
SEAL (Ours)	1/4	39.3	40.9	38.6	31.0	41.1	34.8	33.2

Table 2. **Ablation studies of prediction with partially-observed videos on LVBench.** When only partial videos are visible, the performance of traditional uniform sampling drops significantly, while our method shows more reasonable results.

Ablation on Different α. Figure 1 shows the performance trends across various categories for different values of α. $\alpha = 0.9$ achieves the best overall trade-off, reaching peak with the highest overall accuracy of **45.9**. Conversely, extreme values like $\alpha = 0.0$ or 1.0 lead to declines in several metrics, highlighting that both diversity and relevance are essential. Therefore, $\alpha = 0.9$ is the optimal choice for experiments, delivering peak performance and a well-rounded balance across all categories.

Effectiveness of Encoder for Relevance. Table 3 presents the relevance results computed using the BLIP (Base), CLIP (ViT-L/14) and the BLIP2 (Large) models. The results demonstrate that stronger models achieve higher effectiveness in computing relevance scores, leading to significant performance gains for SEAL.

Model	Overall	KIR	EU	Sum	ER	Rea	TG
SEAL w/ BLIP2	45.9	51.5	41.3	39.7	47.9	43.3	32.3
SEAL w/ CLIP	42.9	48.5	38.6	36.2	46.2	36.3	32.7
SEAL w/ BLIP	40.5	41.9	38.6	39.7	40.5	47.2	32.7

Table 3. Comparison of BLIP2 with other methods on LVBench.

B. Additional Results and Discussions

Comparison with LVU methods on LVBench. We provide additional comparison with LVU methods on LVBench in Table 4. For a fair comparison, we follow those methods to use a 7B LLM. SEAL maintains superior performance with a much smaller LLM (7B), demonstrating the effectiveness of our proposed method.

Model	Overall	KIR	EU	Sum	ER	Rea	TG
MovieChat [9]	22.5	25.9	23.1	17.2	21.3	24.0	22.3
TimeChat [8]	22.3	25.9	21.7	24.1	21.9	25.0	22.7
MA-LLM [4]	24.5	25.4	25.8	22.4	22.3	26.9	21.8
SEAL (7B)	36.6	44.3	33.7	27.6	36.9	32.8	30.9

Table 4. Comparison with other long video representations on LVBench.

Number of different tokens. The subset of tokens is learned as an optimization problem in Section 3.2, and the composition of tokens varies on different inputs and tasks. The averaged percentages of scene, object, and action tokens are 62.5%, 26.1%, 11.4% on LVBench, 54.3%, 25.6%, 20.1% on Moviechat, 88.5% scene tokens and 11.5% action tokens on Ego4d-NLQ. Since Ego4D-NLQ is a temporal localization task, we only utilize scene and action tokens.

Result Analysis. We evaluated the distribution of answers generated by different models, following [11], as shown in Figure 2. The Ground-Truth exhibits a fairly balanced distribution among A, B, C, and D, indicating a well-distributed dataset where no single category is dispropor-

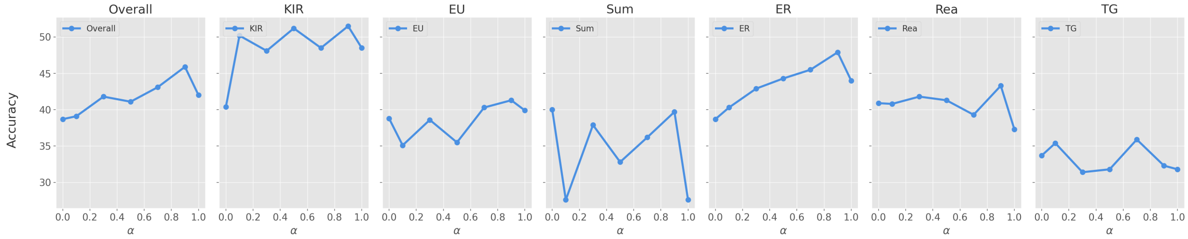


Figure 1. **Ablation studies of different values of α on LVBench.** $\alpha = 0.9$ achieves the best performance across different tasks except for temporal grounding (TG).

tionately represented. MovieChat and LWM shows a dominance of category A, with significantly smaller contributions from other categories, suggesting a lack of diversity in predictions. In Gemini 1.5 Pro, the "Others" category is significantly high, indicating that Gemini 1.5 Pro produces a notable number of unrelated outputs. Our method demonstrates a distribution close to Ground-Truth, showing strong generalization and robustness.

We evaluate performance across various video categories in Table 5. The Human benchmark achieves the highest accuracy across all categories, with an overall accuracy of 94.4%. Ours method achieves an overall accuracy of 45.9%, representing a clear improvement over InternVL2-40B and Qwen2-VL-72B. This demonstrates the our model's ability to generalize better across different video categories. However, the performance in the Cartoon category shows less improvement relative to other categories, indicating potential challenges in handling stylized or abstract visual content. While our method shows clear improvements over existing models, there remains a substantial gap with the Human benchmark across all categories. Further study is needed to enhance the model's understanding of long videos.

C. Additional Qualitative Results

Figure 3 presents additional qualitative results of SEAL on LVBench, showcasing its ability to focus on relevant semantic tokens and provide correct answers to various types of questions. Compared to InterVL2-40B, SEAL effectively attends to critical entities, such as "tattoo" and "man's arm" (Q3.a), distinct "rain forest plants" and "rain forest leaves" (Q3.b), "tall hat woman", "dog", and the "performing" activity (Q4.a), as well as the "black and white dog" and its activity (Q4.b), resulting in accurate answers. In contrast, the answers provided by InterVL2-40B are C, D, C, and C for Q3.a, Q3.b, Q4.a, and Q4.b, respectively. This highlights that InterVL2-40B struggles to capture key information, such as "tattoo", "tall hat woman", and to distinguish "rain forest" from "forest" (InterVL2-40B chose "forest" failing to capture subtle features related to "rain forest"), as well as critical details about the main characters and activities in the scene. These results underscore the superior reasoning capabilities of SEAL.

D. Additional Implementation Details

D.1. Token Extraction

Scene token. We use the full frames to represent scene tokens. The full frames or clips are fed into encoders (2D or 3D CNN/ViT) to extract the token embeddings. For MovieChat and LVBench, we use a frame-based 2D encoder [3] and [2]. For the Ego4D-NLP dataset, we follow [6] and use a 3D clip-based encoder [5] for processing 23-frame clips.

Object token. For object tokens, we extract masks using from SAM2 [7] *Automatic Mask Generator*. For mask prediction, we sample 64×64 points per image for dense and uniform coverage, with a batch size of 128 points to balance computational efficiency and memory usage. Predicted masks are filtered using a quality threshold of `pred_iou_thresh=0.88`, retaining only masks with high predicted IoU scores, and a stability score threshold of `stability_score_thresh=0.92`, ensuring the robustness of masks under varying binarization cutoffs. To calculate the stability score, the cutoff is shifted by `stability_score_offset=0.99`. Non-maximal suppression (NMS) is applied with an IoU threshold of `box_nms_thresh=0.7` to remove redundant masks. We do not employ additional cropping layers (`crop_n_layers=0`). We extract features based on the mask's bounding box, expand it by 2x to include additional contextual information, and use the same encoder as the scene token for different datasets. We set $N_{\text{key}} = 128$ for MovieChat and $N_{\text{key}} = 64$ for Ego4D-NLP and LVBench datasets.

Action token. For the Ego4D-NLP and LVBench datasets, we use YOLOv10-X [10] with BoT-SORT [1] for extracting action tracklets. For MovieChat, we employ NetTrack [12] for action tracklets. We set $L_{\text{min}} = 8$ and $L_{\text{max}} = 16$ for MovieChat, while for Ego4D-NLP, we set $L_{\text{min}} = 16$ and $L_{\text{max}} = 32$. For LVBench, since the action token encoder [2] is a frame-based encoder, we use the middle frame

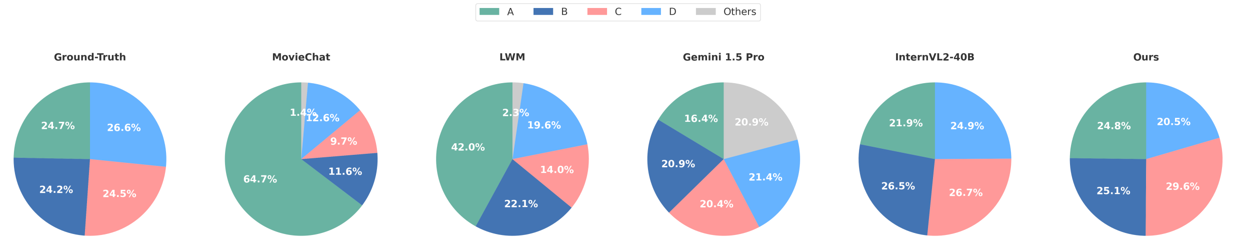


Figure 2. **Distribution of answers generated by different models.** The answers from InterVL2-40B and our method are the closest to the ground truth distribution.

Model	Sports	Documentary	Event Record	Lifestyle	TV Show	Cartoon	Overall
Random predictions	27.5	25.4	23.3	23.3	25.6	25.8	25.1
Random tokens	25.4	25.9	25.6	26.2	24.4	21.6	24.8
Human	96.3	89.8	87.4	98.4	97.2	95.8	94.4
InternVL2-40B	43.5	45.2	38.9	41.6	32.8	36.4	39.5
Qwen2-VL-72B	43.0	42.6	40.8	41.0	42.0	38.9	41.3
SEAL (Ours)	49.2	49.2	48.1	46.7	44.4	39.2	45.9

Table 5. **Evaluation across different video categories on LVBench.** Comparing our method with baselines and state-of-the-art approaches on different video categories. Our method consistently outperforms state-of-the-art approaches on all categories. Although our method has made substantial improvements over lower-bound baselines (Random tokens and Random predictions), it still has a significant gap compared with the upper-bound baseline of human performance.

of all action tracklets as the action token candidates.

In Attention Learning stage, we sample in total 256 tokens for MovieChat, 200 / 450 tokens for Ego4D-NLP and 16 tokens for LVBench. Note that since the task of Ego4D-NLP is temporal grounding, we only used action tokens and scene tokens to ensure temporal consistency.

D.2. LLM Heads and LLM-based Evaluation

For the MovieChat dataset, we provide the large language model with the following prompt for the Video QA task:

```
"You are able to understand
the visual content that the
user provides.  Follow the
instructions carefully and
explain your answers."
```

For the LVBench dataset, given a question and options, we use the prompt for the Video QA multiple choice task:

```
"Please select the best answer
from the options above and
directly provide the letter
representing your choice without
giving any explanation."
```

Following [9], we use LLM-Assisted Evaluation for the video question-answering task when evaluating MovieChat dataset. Given the question, the correct answer, and the predicted answer provided by different methods, the LLM assistants should return a True or False judgment along with a relative score ranging from 0 to 5. we provide the large language model with the following prompt:

```
"Provide your evaluation
only as a yes/no and score
where the score is an integer
value between 0 and 5, with
5 indicating the highest
meaningful match."
```

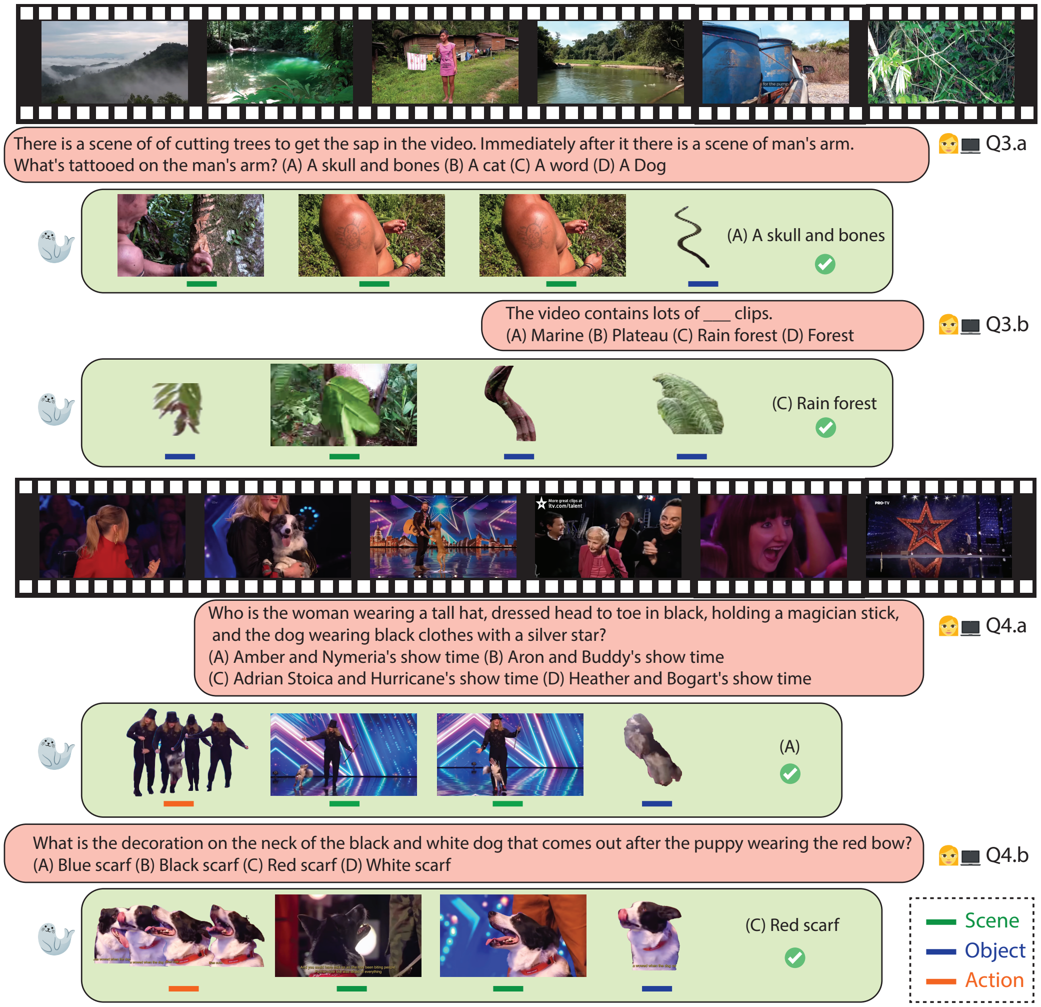
References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *ArXiv*, abs/2206.14651, 2022. 2

[2] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 2

[3] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, 2023. 2

[4] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim.


There is a scene of of cutting trees to get the sap in the video. Immediately after it there is a scene of man's arm. What's tattooed on the man's arm? (A) A skull and bones (B) A cat (C) A word (D) A Dog

The video contains lots of ___ clips. (A) Marine (B) Plateau (C) Rain forest (D) Forest

Who is the woman wearing a tall hat, dressed head to toe in black, holding a magician stick, and the dog wearing black clothes with a silver star? (A) Amber and Nymeria's show time (B) Aron and Buddy's show time (C) Adrian Stoica and Hurricane's show time (D) Heather and Bogart's show time

What is the decoration on the neck of the black and white dog that comes out after the puppy wearing the red bow? (A) Blue scarf (B) Black scarf (C) Red scarf (D) White scarf

Figure 3. **Additional qualitative results on LVBench**.SEAL attends to relevant entities such as "tattoo" and "man's arm" (Q3.a), different "rain forest plants" and "rain forest leaves" (Q3.b), "tall hat woman", "dog", and "performing" activity (Q4.a), " black and white dog" and its activity (Q4.b) and correctly answers these questions. However, the answers provided by InterVL2-40B are C, D, C, C for Q3.a, Q3.b, Q4.a, and Q4.b, respectively. This indicates that InterVL2-40B fails to capture key information such as "tattoo", "rainforest", and important details about the main characters in the performance.

Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *CVPR*, 2024. 1

[5] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *arXiv preprint arXiv:2206.01670*, 2022. 2

[6] Fangzhou Mu, Sicheng Mo, and Yin Li. Snag: Scalable and accurate video grounding. In *CVPR*, pages 18930–18940, 2024. 2

[7] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos, 2024. 2

[8] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *CVPR*, pages 14313–14323, 2024. 1

[9] E. Song, W. Chai, G. Wang, Y. Zhang, H. Zhou, F. Wu, X. Guo, T. Ye, Y. Lu, JN. Hwang, and G. Wang. MovieChat: From dense token to sparse memory for long video understanding. In *CVPR*, 2024. 1, 3

[10] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yolov10: Real-time end-to-end object detection. *ArXiv*, abs/2405.14458, 2024. 2

[11] W. Wang, Z. He, W. Hong, Y. Cheng, X. Zhang, J. Qi, S. Huang, B. Xu, Y. Dong, M. Ding, and J. Tang. LVBench:

An extreme long video understanding benchmark, 2024. 1

[12] Guangze Zheng, Shijie Lin, Haobo Zuo, Changhong Fu, and Jia Pan. NetTrack: Tracking Highly Dynamic Objects with a Net. In *CVPR*, 2024. 2