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-

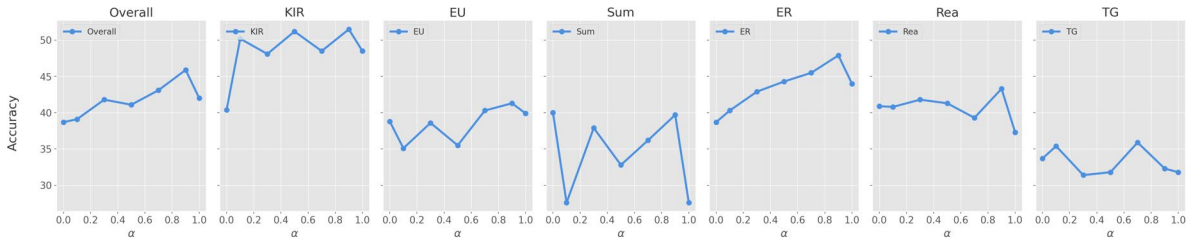

Figure 1. **Ablation studies of different values of α on LVBench.** $\alpha = 0.9$ achieves the best performance across different tasks except for temporal grounding (TG).

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

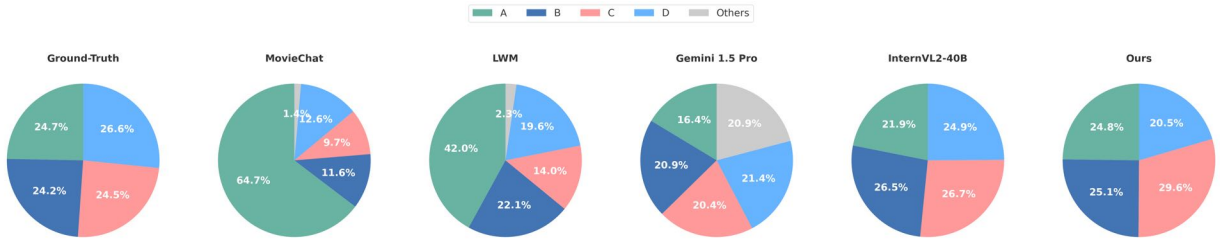

Figure 2. **Distribution of answers generated by different models.** The answers from InterVL2-40B and our method are the closest to the ground truth distribution.

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.

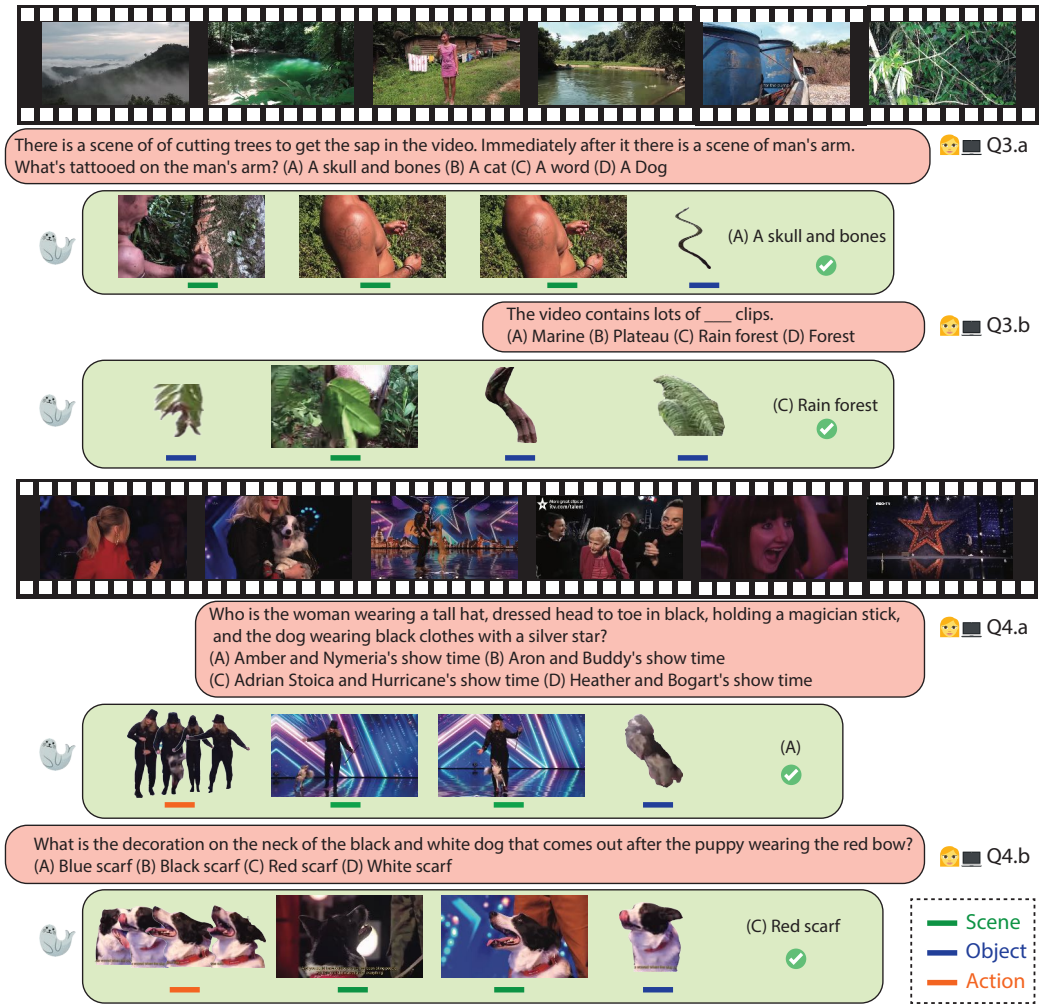

Figure 3. **Additional qualitative results on LVBench**.SEAL attends to relevant entities such as "tattoo" and "man's arm" (Q3.a), different "rain forest plants" and "rain forest leaves" (Q3.b), "tall hat woman", "dog", and "performing" activity (Q4.a), " black and white dog" and its activity (Q4.b) 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