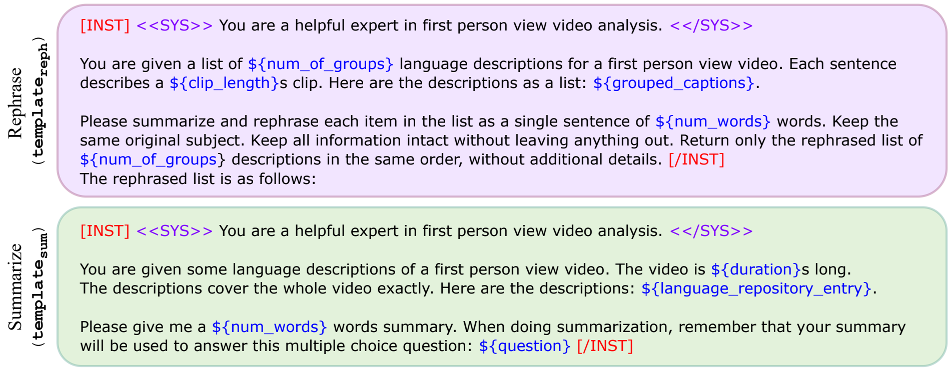
Language Repository for Long Video Understanding - Supplementary

Kumara Kahatapitiya   Kanchana Ranasinghe   Jongwoo Park   Michael S. Ryoo
Stony Brook University
kkahatapitiy@cs.stonybrook.edu

A Appendix



Rephrase ($\texttt{template}_{\texttt{reph}}$)

[INST] <<SYS>> You are a helpful expert in first person view video analysis. <</SYS>>

You are given a list of ${num_of_groups} language descriptions for a first person view video. Each sentence describes a ${clip_length}s clip. Here are the descriptions as a list: ${grouped_captions}.

Please summarize and rephrase each item in the list as a single sentence of ${num_words} words. Keep the same original subject. Keep all information intact without leaving anything out. Return only the rephrased list of ${num_of_groups} descriptions in the same order, without additional details. [/INST]
The rephrased list is as follows:

Summarize ($\texttt{template}_{\texttt{sum}}$)

[INST] <<SYS>> You are a helpful expert in first person view video analysis. <</SYS>>

You are given some language descriptions of a first person view video. The video is ${duration}s long. The descriptions cover the whole video exactly. Here are the descriptions: ${language_repository_entry}.

Please give me a ${num_words} words summary. When doing summarization, remember that your summary will be used to answer this multiple choice question: ${question} [/INST]

Figure A.1: LLM prompt templates in LangRepo: Here, we show the zero-shot prompt templates used for rephrasing ($\texttt{template}_{\texttt{reph}}$) and summarizing ($\texttt{template}_{\texttt{sum}}$) operations. Rephrase prompt needs a list of grouped captions as input, while its output adheres to more-strict requirements (*e.g.* same order, #items) needed for correct parsing. Summarize prompt takes in each repository entry and generates a more-flexible (*i.e.*, open-ended) output, optionally conditioning on query.

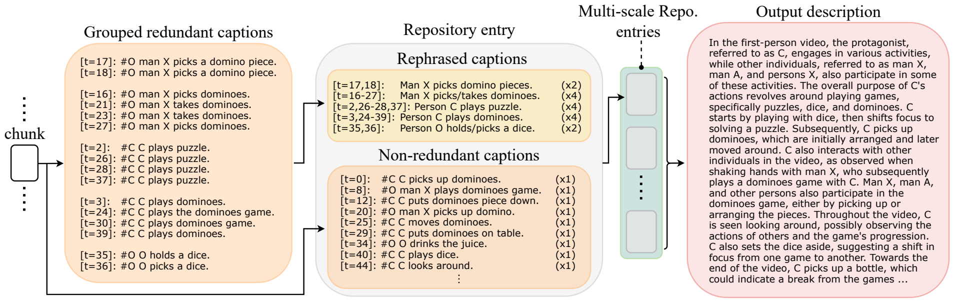
A.1 Rephrasing and Summarizing Templates

In LangRepo, during our write operation, we wrap the grouped captions in a rephrasing-template ($\texttt{template}_{\texttt{reph}}$) before calling the LLM. This template is shown in Fig. A.1 (top). Here, we provide a few constraints so that we can parse the rephrased output correctly. For instance, (1) we want rephrasing to happen within each group (*i.e.*, not across groups), and (2) we expect the output to have the same number of groups in the same order. Similarly, during our read operation, we wrap repository entries in a summarizing-template ($\texttt{template}_{\texttt{sum}}$) as shown in Fig. A.1 (bottom). This template follows a relatively simpler formulation, having a word-limit and query-conditioning.

A.2 Prompting for VQA

As the evaluation setup, we consider multiple-choice visual question-answering (VQA) on long videos. Given the close-ended answer formulation, we can consider two different classifiers to make the prediction: (1) a Generative classifier, which directly generates the answer choice, or (2) a Log-likelihood classifier, which select the most-probable choice based on the joint-probability of tokens in each answer option given the description and the question. The latter generally performs better, as it is less-prone to hallucinations (*i.e.*, prediction is explicitly constrained to answer choices). However, it is also sensitive to the prompts we use. Hence, we include a discussion on prompting in the following subsections.

Workshop on Video-Language Models at 38th Conference on Neural Information Processing Systems (NeurIPS 2024).



```
[t=17]: #O man X picks a domino piece.
[t=18]: #O man X picks a domino piece.

[t=16]: #O man X picks dominoes.
[t=21]: #O man X takes dominoes.
[t=23]: #O man X takes dominoes.
[t=27]: #O man X picks dominoes.

[t=2]:  #C C plays puzzle.
[t=26]: #C C plays puzzle.
[t=28]: #C C plays puzzle.
[t=37]: #C C plays puzzle.

[t=3]:  #C C plays dominoes.
[t=24]: #C C plays the dominoes game.
[t=30]: #C C plays dominoes game.
[t=39]: #C C plays dominoes.

[t=35]: #O O holds a dice.
[t=36]: #O O picks a dice.
```

```
[t=17,18]:        Man X picks domino pieces.      (x2)
[t=16-27]:        Man X picks/takes dominoes.     (x4)
[t=2,26-28,37]: Person C plays puzzle.            (x4)
[t=3,24-39]:    Person C plays dominoes.          (x4)
[t=35,36]:        Person O holds/picks a dice.    (x2)
```

```
[t=0]:  #C C picks up dominoes.            (x1)
[t=8]:  #O man X plays dominoes game.      (x1)
[t=12]: #C C puts dominoes piece down.     (x1)
[t=20]: #O man X picks up domino.          (x1)
[t=25]: #C C moves dominoes.               (x1)
[t=29]: #C C puts dominoes on table.       (x1)
[t=34]: #O O drinks the juice.             (x1)
[t=40]: #C C plays dice.                   (x1)
[t=44]: #C C looks around.                 (x1)
                    ⋮
```

In the first-person video, the protagonist, referred to as C, engages in various activities, while other individuals, referred to as man X, man A, and persons X, also participate in some of these activities. The overall purpose of C's actions revolves around playing games, specifically puzzles, dice, and dominoes. C starts by playing with dice, then shifts focus to solving a puzzle. Subsequently, C picks up dominoes, which are initially arranged and later moved around. C also interacts with other individuals in the video, as observed when shaking hands with man X, who subsequently plays a dominoes game with C. Man X, man A, and other persons also participate in the dominoes game, either by picking up or arranging the pieces. Throughout the video, C is seen looking around, possibly observing the actions of others and the game's progression. C also sets the dice aside, suggesting a shift in focus from one game to another. Towards the end of the video, C picks up a bottle, which could indicate a break from the games ...

Figure A.2: A qualitative example of a single-iteration LangRepo entry: Given a video chunk, redundant captions are grouped together and rephrased to be more-concisely written to the repository, along with additional metadata. Other non-redundant captions are written directly. Next, such repository entries are summarized to generate output descriptions when reading.

Generative classifier: Here, we direcly prompt the LLM to generate the correct answer, conditioned on the descriptions generated by LangRepo, the question and the answer options (inspired by [14]). To make sure that the output can be parsed, we provide additional guiding instructions and any syntax specific to the LLM (Mistral [2]). This also discourages any hallucinations. On all benchmarks, we use the common prompt given below.

> "[INST] «SYS» You are a helpful expert in first person view video analysis. «/SYS» Please provide a single-letter answer (A, B, C, D, E) to the following multiple-choice question, and your answer must be one of the letters (A, B, C, D, or E). You must not provide any other response or explanation. You are given some language descriptions of a first person view video. The video is ${duration} seconds long. Here are the descriptions: ${description}.\n You are going to answer a multiple choice question based on the descriptions, and your answer should be a single letter chosen from the choices.\n Here is the question: ${question}.\n Here are the choices.\n A: ${optionA}\n B: ${optionB}\n C: ${optionC}\n D: ${optionD}\n E: ${optionE}\n [/INST]"

Log-likelihood classifier: Inspired by [8], in this setup, we prompt the LLM with each answer option separately, and select the highest-probable answer. The probability is computed only on the tokens of the answer option, conditioned on the input sequence. In our experiments, we notice that the effectiveness of this method is sensitive to the prompt. This is due to the question-answer formats in the dataset considered. For instance, EgoSchema [6] consists of full-sentence answers, whereas NExT-QA [13] consists of answer phrases. Hence, the latter benefits from additional guidance from formatting within the prompt template. More specifically, on EgoSchema [6], our prompt has the following format.

"${description} ${question} ${answer_option}"

Here, the probability is computed only on ${answer_option}. However, on the benchmarks based on NExT-QA [13] data, our prompt has the following format with more structure.

> "${description} Based on the description above, answer the following question: ${question}? Select one of these choices as the answer:\n A: ${optionA}\n B: ${optionB}\n C: ${optionC}\n D: ${optionD}\n E: ${optionE}\n The correct answer is, ${option_id}: ${answer_option}"

Here, the probability is computed only on ${option_id}:${answer_option}. We observe that neither prompt template works as effective when interchanged.

A.3 Qualitative examples of repository entries

We present qualitative examples from EgoSchema [6] dataset to better clarify the operations in LangRepo. In Fig. A.2, we show the format of repository entries in a single iteration. Here, non-redundant captions from the input get directly written to the repo. In contrast, any redundant captions—grouped based on similarity— get rephrased as concise descriptions (one description per-group). Each repository description may come with additional metadata such as timestamps and #occurrences to avoid any loss of meaningful information due to pruning.

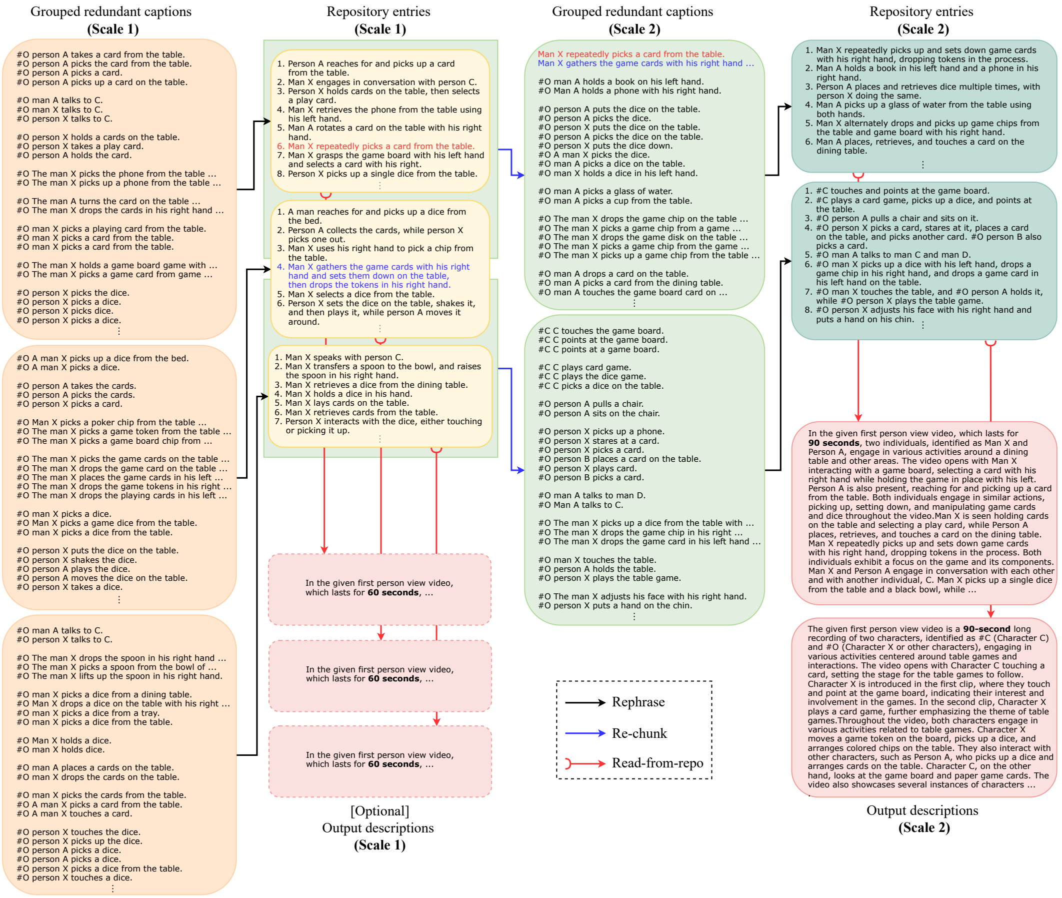

Grouped redundant captions
(Scale 1)
#O person A takes a card from the table.
#O person A picks the card from the table.
#O person A picks a card.
#O person A picks up a card on the table.
#O man A talks to C.
#O man X talks to C.
#O person X talks to C.
#O person X holds a cards on the table.
#O person X takes a play card.
#O person A holds the card.
#O The man X picks the phone from the table ...
#O The man X picks up a phone from the table ...
#O The man A turns the card on the table ...
#O The man X drops the cards in his right hand ...
#O man X picks a playing card from the table.
#O man X picks a card from the table.
#O man X picks a card from the table.
#O The man X holds a game board game with ...
#O The man X picks a game card from game ...
#O person X picks the dice.
#O person X picks a dice.
#O person X picks dice.
#O person X picks a dice.
⋮
#O A man X picks up a dice from the bed.
#O A man X picks a dice.
#O person A takes the cards.
#O person A picks the cards.
#O person X picks a card.
#O Man X picks a poker chip from the table ...
#O The man X picks a game token from the table ...
#O The man X picks a game board chip from ...
#O The man X picks the game cards on the table ...
#O The man X drops the game card on the table ...
#O The man X places the game cards in his left ...
#O The man X drops the game tokens in his right ...
#O The man X drops the playing cards in his left ...
#O man X picks a dice.
#O Man X picks a game dice from the table.
#O man X picks a dice from the table.
#O person X puts the dice on the table.
#O person X shakes the dice.
#O person A plays the dice.
#O person A moves the dice on the table.
#O person X takes a dice.
⋮
#O man A talks to C.
#O person X talks to C.
#O The man X drops the spoon in his right hand ...
#O The man X picks a spoon from the bowl of ...
#O The man X lifts up the spoon in his right hand.
#O man X picks a dice from a dining table.
#O Man X drops a dice on the table with his right ...
#O man X picks a dice from a tray.
#O man X picks a dice from the table.
#O Man X holds a dice.
#O man X holds dice.
#O man A places a cards on the table.
#O man X drops the cards on the table.
#O man X picks the cards from the table.
#O A man X picks a card from the table.
#O A man X touches a card.
#O person X touches the dice.
#O person X picks up the dice.
#O person A picks a dice.
#O person A picks a dice.
#O person X picks a dice from the table.
#O person X touches a dice.
⋮
Repository entries
(Scale 1)
1. Person A reaches for and picks up a card from the table.
2. Man X engages in conversation with person C.
3. Person X holds cards on the table, then selects a play card.
4. Man X retrieves the phone from the table using his left hand.
5. Man A rotates a card on the table with his right hand.
6. Man X repeatedly picks a card from the table.
7. Man X grasps the game board with his left hand and selects a card with his right.
8. Person X picks up a single dice from the table.
⋮
1. A man reaches for and picks up a dice from the bed.
2. Person A collects the cards, while person X picks one out.
3. Man X uses his right hand to pick a chip from the table.
4. Man X gathers the game cards with his right hand and sets them down on the table, then drops the tokens in his right hand.
5. Man X selects a dice from the table.
6. Person X sets the dice on the table, shakes it, and then plays it, while person A moves it around.
⋮
1. Man X speaks with person C.
2. Man X transfers a spoon to the bowl, and raises the spoon in his right hand.
3. Man X retrieves a dice from the dining table.
4. Man X holds a dice in his hand.
5. Man X lays cards on the table.
6. Man X retrieves cards from the table.
7. Person X interacts with the dice, either touching or picking it up.
⋮
In the given first person view video, which lasts for **60 seconds**, ...
In the given first person view video, which lasts for **60 seconds**, ...
In the given first person view video, which lasts for **60 seconds**, ...
[Optional]
Output descriptions
(Scale 1)
Grouped redundant captions
(Scale 2)
Man X repeatedly picks a card from the table.
Man X gathers the game cards with his right hand ...
#O man A holds a book on his left hand.
#O Man A holds a phone with his right hand.
#O person A puts the dice on the table.
#O person A picks the dice.
#O person X puts the dice on the table.
#O person A picks the dice on the table.
#O person X puts the dice down.
#O A man X picks the dice.
#O man A picks a dice on the table.
#O man X holds a dice in his left hand.
#O man A picks a glass of water.
#O man A picks a cup from the table.
#O The man X drops the game chip on the table ...
#O The man X picks a game chip from a game ...
#O The man X drops the game disk on the table ...
#O The man X picks a game chip from the game ...
#O The man X picks up a game chip from the table ...
#O man A drops a card on the table.
#O man A picks a card from the dining table.
#O man A touches the game board card on ...
⋮
#C C touches the game board.
#C C points at the game board.
#C C points at a game board.
#C C plays card game.
#C C plays the dice game.
#C C picks a dice on the table.
#O person A pulls a chair.
#O person A sits on the chair.
#O person X picks up a phone.
#O person X stares at a card.
#O person X picks a card.
#O person B places a card on the table.
#O person X plays card.
#O person B picks a card.
#O man A talks to man D.
#O Man A talks to C.
#O The man X picks up a dice from the table with ...
#O The man X drops the game chip in his right ...
#O The man X drops the game card in his left hand ...
#O man X touches the table.
#O person A holds the table.
#O person X plays the table game.
#O The man X adjusts his face with his right hand.
#O person X puts a hand on the chin.
⋮
Rephrase
Re-chunk
Read-from-repo
Repository entries
(Scale 2)
1. Man X repeatedly picks up and sets down game cards with his right hand, dropping tokens in the process.
2. Man A holds a book in his left hand and a phone in his right hand.
3. Person A places and retrieves dice multiple times, with person X doing the same.
4. Man A picks up a glass of water from the table using both hands.
5. Man X alternately drops and picks up game chips from the table and game board with his right hand.
6. Man A places, retrieves, and touches a card on the dining table.
⋮
1. #C touches and points at the game board.
2. #C plays a card game, picks up a dice, and points at the table.
3. #O person A pulls a chair and sits on it.
4. #O person X picks a card, stares at it, places a card on the table, and picks another card. #O person B also picks a card.
5. #O man A talks to man C and man D.
6. #O man X picks up a dice with his left hand, drops a game chip in his right hand, and drops a game card in his left hand on the table.
7. #O man X touches the table, and #O person A holds it, while #O person X plays the table game.
8. #O person X adjusts his face with his right hand and puts a hand on his chin. ⋮
In the given first person view video, which lasts for **90 seconds**, two individuals, identified as Man X and Person A, engage in various activities around a dining table and other areas. The video opens with Man X interacting with a game board, selecting a card with his right hand while holding the game in place with his left. Person A is also present, reaching for and picking up a card from the table. Both individuals engage in similar actions, picking up, setting down, and manipulating game cards and dice throughout the video.Man X is seen holding cards on the table and selecting a play card, while Person A places, retrieves, and touches a card on the dining table. Man X repeatedly picks up and sets down game cards with his right hand, dropping tokens in the process. Both individuals exhibit a focus on the game and its components. Man X and Person A engage in conversation with each other and with another individual, C. Man X picks up a single dice from the table and a black bowl, while ...
The given first person view video is a **90-second** long recording of two characters, identified as #C (Character C) and #O (Character X or other characters), engaging in various activities centered around table games and interactions. The video opens with Character C touching a card, setting the stage for the table games to follow. Character X is introduced in the first clip, where they touch and point at the game board, indicating their interest and involvement in the games. In the second clip, Character X plays a card game, further emphasizing the theme of table games.Throughout the video, both characters engage in various activities related to table games. Character X moves a game token on the board, picks up a dice, and arranges colored chips on the table. They also interact with other characters, such as Person A, who picks up a dice and arranges cards on the table. Character C, on the other hand, looks at the game board and paper game cards. The video also showcases several instances of characters ...
Output descriptions
(Scale 2)

Figure A.3: A qualitative example of iterative writing and multi-scale reading in LangRepo: Here, we present an example with 2-scales, given captions of a 180s long video. In scale-1, we consider 3 chunks of 60s each, and in scale-2, we `re-chunk` them into 2 chunks of 90s each. We only show the redundant captions that go through pruning, and also, omit any metadata (*e.g.* timestamps) within the repository. In each scale, captions grouped based on similarity get rephrased concisely. To generate inputs of the subsequent scale, we simply order previous repository descriptions in time, and split (*i.e.*, `re-chunk`) into fewer (and, longer) chunks. When reading, each entry in each scale is summarized separately to create output descriptions of various temporal spans. In general, we always consider the last-scale descriptions to be mandatory, but any prior-scale to be optional. Best-viewed with zoom-in.

In Fig. A.3, we further elaborate on multiple iterations (*i.e.*, temporal scales) within the repository, which are generated by iteratively processing increasingly-longer chunks (created by `re-chunk` operation). Such scales remove redundancies across different temporal spans and generate more high-level information. For instance, in the figure, scale-1 captures a 60s window, while scale-2 observes a 90s window. In scale-1 repository, {*picking-up card*} is in a separate entry from {*playing dice, picking-up chip*}. Such details when combined can provide high-level information (*e.g.* figuring out it is *dice poker*, instead of just a *card game* or a *dice game*). In output descriptions in scale-2 we observe reasoning over longer temporal context, which is not present in scale-1. During reading, we can decide to summarize information at various temporal scales to generate output descriptions.

A.4 Design decisions

Choice of backbone LLM and Text encoder: We consider different open-source LLMs, namely, LLama2 [11] and Mistral [2]. We ablate the choice of LLM within LLoVi [14] framework, as shown in Table A.1a. We observe that Mistral-7B is better at video reasoning compared to LLama2-13B. When identifying redundancies, we use a similarity-based grouping with the help of text embeddings. This gives more control on what to prune and how much to prune, compared to a direct LLM-call for pruning. Table A.1b Shows that CLIP [7] works better than Sentence-T5 [9] in this regard.

Table A.1: Ablating design decisions on EgoSchema [6]: We evaluate different design decisions of our framework on EgoSchema 500-video subset for zero-shot video VQA.

(a) **Choice of LLM**: With framework in [14], Mistral [2] significantly outperforms LLama2 [11] even at a smaller scale.

LLM	Scale	Acc.
Llama2 [11]	13B	43.0
Mistral [2]	7B	50.8

(b) **Text encoder**: CLIP [7] outperforms Sentence-T5 [9] (trained with setntence objective) for similarity-based pruning.

Text encoder	Acc.
Sentence-T5-XL [9]	56.4
CLIP-L/14 [7]	57.8

(c) **VQA classifier**: Log-likelihood classifier outperforms generative classifier for close-ended VQA.

VQA classifier	Acc.
Genearative	57.8
Log-likelihood	60.8

(d) **Video input**: Feeding short captions chunk-by-chunk to the LLM is empirically-better than feeding all-at-once.

Streaming setup	Acc.
LLoVi [14]	50.8
Chunk-based LLoVi	57.8
LangRepo (ours)	60.8

(e) **Repository setup**: In LangRepo, more iterations with finer chunks during writing, and multiple scales during reading improves VQA performance.

#Iter	#Ch	Scales-read	Acc.
1	[2]	1	57.0
1	[4]	1	60.8
3	[4,3,2]	1	58.4
3	[4,3,2]	2	59.4
3	[4,3,2]	3	61.2

(f) **Metadata within repository**: Timesteps do not help, yet #occurrences help weigh repo-entries properly.

Model	Acc.
LangRepo (ours)	60.8
+ tstmp	60.4
+ occ	61.4
+ tstmp + occ	58.2

(g) **Captioner**: Clip-level captions outperform frame-level captions. Gap to oracle is still significant.

Captions	Acc.
BLIP-2 [4]	55.4
LLaVA-1.5 [5]	58.4
LaViLa [15]	60.8
Oracle	69.2

Classifier for close-ended VQA: As discussed in Sec. A.2, we compare log-likelihood based classifier with a generative classifier. Among these, we find the former to be better-performing as shown in Table A.1c. Yet, it is more-sensitive to the prompt template.

Processing videos as chunks: Our decision to consume longer videos as chunks is motivated by prior work [12, 10]. It allows us to not lose short-term details, while also keeping track of long-term dependencies via multi-scale processing. This choice is validated by the results in Table A.1d. Additionally, although not explored in the scope of this paper, such a setup integrates well with temporally-fine-grained prediction tasks, where an LLM makes multiple predictions over time.

Repository setup: In the formulation of LangRepo we ablate different hyperparameter settings related to the number of repo-updates (#iterations), the number of video chunks in each iteration (#chunks), and multiple temporal-scales considered when reading data in repository. In Table A.1e, we make two observations: (1) more update iterations with finer chunks (higher #chunks per iteration) can preserve more-useful information, and (2) reading information in multiple temporal-scales is consistently better. In terms of metadata that we preserve, we see in Table A.1f that #occurrences do help when summarizing (by weighting each entry accordingly), yet timestamps do not provide meaningful improvement.

Captioner quality: In Table A.1g, we evaluate the quality of captions consumed by LangRepo. By default, we use short-clip captions from LaViLa [15], which outperform frame-level captions (BLIP-2 [4], LLaVA-1.5 [5]). Oracle captions from Ego4D show the performance upper-bound.

A.5 Dataset details

EgoSchema: EgoSchema [6] is a long-video VQA dataset derived from Ego4D [1], by semi-automatically generating and verifying mutliple-choice questions. Each video is 3 minutes long, and the questions are filtered to have a long temporal certificate (i.e., at least 100 seconds of a video is required to answer the question). The public validation subset consists of 500 videos, each with 5 answer-choices and the correct ground-truth answer. The fullset of 5K videos is hosted as a Kaggle challenge. There is no training split for this dataset, motivating zero-shot evaluation.

NExT-QA: NExT-QA [13] is a popular VQA dataset with videos up to 2 minutes long, at an average of 44 seconds. It consists of 52k open-ended questions and 48k close-ended questions (i.e., multiple-choice with 5 answer options). We consider zero-shot evaluation on 5k validation set.

IntentQA: IntentQA [3] is based on the same NExT-QA videos (specifically, the ones for temporal and causal reasoning), yet re-purposed focusing on intent-related questions. A new set of 16k multiple-choice questions are provided for 4.3k videos, each with 5 answer choices. In our evaluation, we focus on zero-shot setting on test set of 2k questions.

References

[1] 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.

[2] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[3] Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11963–11974, 2023.

[4] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[5] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[6] 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, 2024.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

[8] Kanchana Ranasinghe, Xiang Li, Kumara Kahatapitiya, and Michael S Ryoo. Understanding long videos in one multimodal language model pass. *arXiv preprint arXiv:2403.16998*, 2024.

[9] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[10] Michael S Ryoo, Keerthana Gopalakrishnan, Kumara Kahatapitiya, Ted Xiao, Kanishka Rao, Austin Stone, Yao Lu, Julian Ibarz, and Anurag Arnab. Token turing machines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19070–19081, 2023.

[11] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[12] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13587–13597, 2022.

[13] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021.

[14] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. *arXiv preprint arXiv:2312.17235*, 2023.

[15] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6586–6597, 2023.