Generating Questions, Answers, and Distractors for Videos: Exploring Semantic Uncertainty of Object Motions

Wenjian Ding[1], **Yao Zhang**[*2], **Jun Wang**[3], **Adam Jatowt**[4], **Zhenglu Yang**[*1]
[1]TMCC, Key Laboratory of DISSec, College of Computer Science, Nankai University
[2]School of Statistics and Data Science, AAIS, Nankai University, Tianjin, China
[3]College of Mathematics and Statistics Science, Ludong University
[4]University of Innsbruck, Austria
wjding@mail.nankai.edu.cn, yaozhang@nankai.edu.cn, junwang@mail.nankai.edu.cn,
adam.jatowt@uibk.ac.at, yangzl@nankai.edu.cn

Abstract

Video Question-Answer-Distractors (QADs) show promising values for assessing the performance of systems in perceiving and comprehending multimedia content. Given the significant cost and labor demands of manual annotation, existing large-scale Video QADs benchmarks are typically generated automatically using video captions. Since video captions are incomplete representations of visual content and susceptible to error propagation, direct generation of QADs from video is crucial. This work first leverages a large vision-language model for video QADs generation. To enhance the consistency and diversity of the generated QADs, we propose utilizing temporal motion to describe the video objects. In addition, we design a selection mechanism that chooses diverse temporal object motions to generate diverse QADs focusing on different objects and interactions, maximizing overall semantic uncertainty for a given video. Evaluation on the NExT-QA and Perception Test benchmarks demonstrates that the proposed approach significantly improves both the consistency and diversity of QADs generated by a range of large vision-language models, thus highlighting its effectiveness and generalizability.

1 Introduction

Video Question-Answer-Distractors (QADs) (Lei et al., 2018; Xiao et al., 2021; Pătrăucean et al., 2023) facilitates the evaluation of video-language understanding across modalities, offering a benchmark for both human and machine performance. For instance, online education videos MOOC requires students to answer questions to assess their understanding of the video materials. Besides, Video QADs serve as established benchmarks for evaluating multimodal models (OpenAI, 2024; Google, 2024; Cheng et al., 2024). The manual generation of Video QADs necessitates the analysis of video content, incurring significant costs and resulting in bias (Goyal et al., 2017). This constraint has driven the development of QADs generation via automated algorithms (Li et al., 2016; Xu et al., 2017).

Two primary approaches currently dominate the automatic generation of QADs for videos. The first approach relies on textual information extracted from video captions (Jang et al., 2017) or transcriptions (Yang et al., 2021), which are susceptible to incomplete representations (Zeng et al., 2017) and error propagation (Su et al., 2021). The second approach generates QADs directly from the video content. Existing research in this domain has primarily focused on leveraging question types (Su et al., 2021) or solely object information (Wang et al., 2020) for question-answer pair generation. However, this crucial aspect of video understanding remains unaddressed by current methodologies. As illustrated in Figure 1, the QADs generated by LLaVA-NeXT cannot effectively assess the understanding of motions and interactions of the videos. In contrast, the QADs generated by our model, *MaxSem*, are specifically focused on the temporal object motions and interactions present within the video.

In this work, we propose ***MaxSem***, a framework to **Max**imize the **Sem**antic uncertainty of generated question-answer-distractors for video, ensuring both consistency and diversity. We introduce temporal object motions, which provide a description of object movements over time within the video. The generated QADs leverage object motions and interactions, which are widely present in the ground truth datasets, thereby enhancing consistency. Furthermore, a selection mechanism is proposed to sample diverse combinations of temporal object motions when generating multiple QADs from a single video, thereby maximizing semantic uncertainty and ensuring QADs diversity. We conduct extensive experiments on the NExT-QA

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2025, pages 7207–7220
July 27 - August 1, 2025

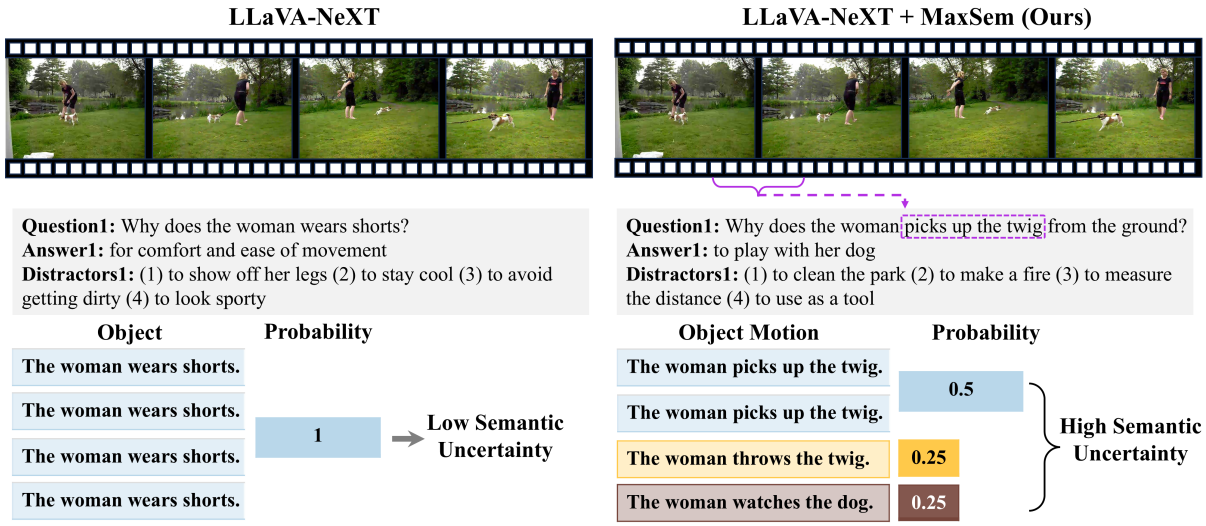

Figure 1: Our method facilitates the generation of QADs focused on object motion with high semantic uncertainty.

and Perception Test benchmarks, empirical results reveal that the proposed method significantly improves both the consistency and diversity of QADs generated by LVLMs. The contribution of this paper lies in the automatic generation of QADs to evaluate video understanding capabilities. Moreover, a prospective direction for future research involves utilizing the generated QADs to improve the performance of existing models.

The main contributions of this work are listed as follows:

- We investigate generating QADs according to videos using LVLMs, leveraging a novel representation of temporal object motions to capture object movements and interactions between different objects over time.
- We introduce a selector mechanism to generate various combinations of temporal object motions, enabling the generation of diverse QADs for a given video, ultimately maximizing semantic uncertainty.
- Extensive experimental results on the NExT-QA and Perception Test benchmarks reveal that the proposed methods can help existing LVLMs to generate consistent and diverse QADs.

2 Related Works

The majority of previous research has concentrated on the generation of a part of or parts of QADs, namely questions, answers, or distractors.

Video Question Generation involves generating meaningful questions from video content or captions. Wang et al. (2020) leveraged video frames, detected objects, and subtitles as input to generate semantically meaningful questions. Su et al. (2021) introduced a Generator-Pretester network that generates question-answer pairs and subsequently validates the generated question by trying to answer it. Guo et al. (2020) extended this work to the multi-turn setting, generating multiple questions based on both dialogue history and video content.

Video Question Answering (Le et al., 2020; Xiao et al., 2022b,a) requires the models to understand both the complex video and language data to correctly generate the answers. Wang et al. (2022) proposed VidIL, a framework that instructs a large language model on video-language tasks based on the temporal-aware template. Li et al. (2022) proposed an invariant grounding framework (IGV) to distinguish the causal scene and emphasize its causal effect on the answer. Li et al. (2023) developed a differentiable selection module that adaptively collects question-critical moments and objects using cross-modal interaction.

Visual Distractor Generation targets to generate challenging distractors according to the visual content. Lu et al. (2022) introduced a reinforcement learning approach to generate distractors from visual images. Luo et al. (2024) proposed leveraging multimodal large language models with Chain-of-Thought reasoning to generate both questions and distractors. Ding et al. (2024a) first proposed to generate questions, answers, and distractors jointly. Ding et al. (2024b) introduced a framework to generate QADs that focus on different regions in the image. However, their work is confined to the analysis of individual images. In this paper, we

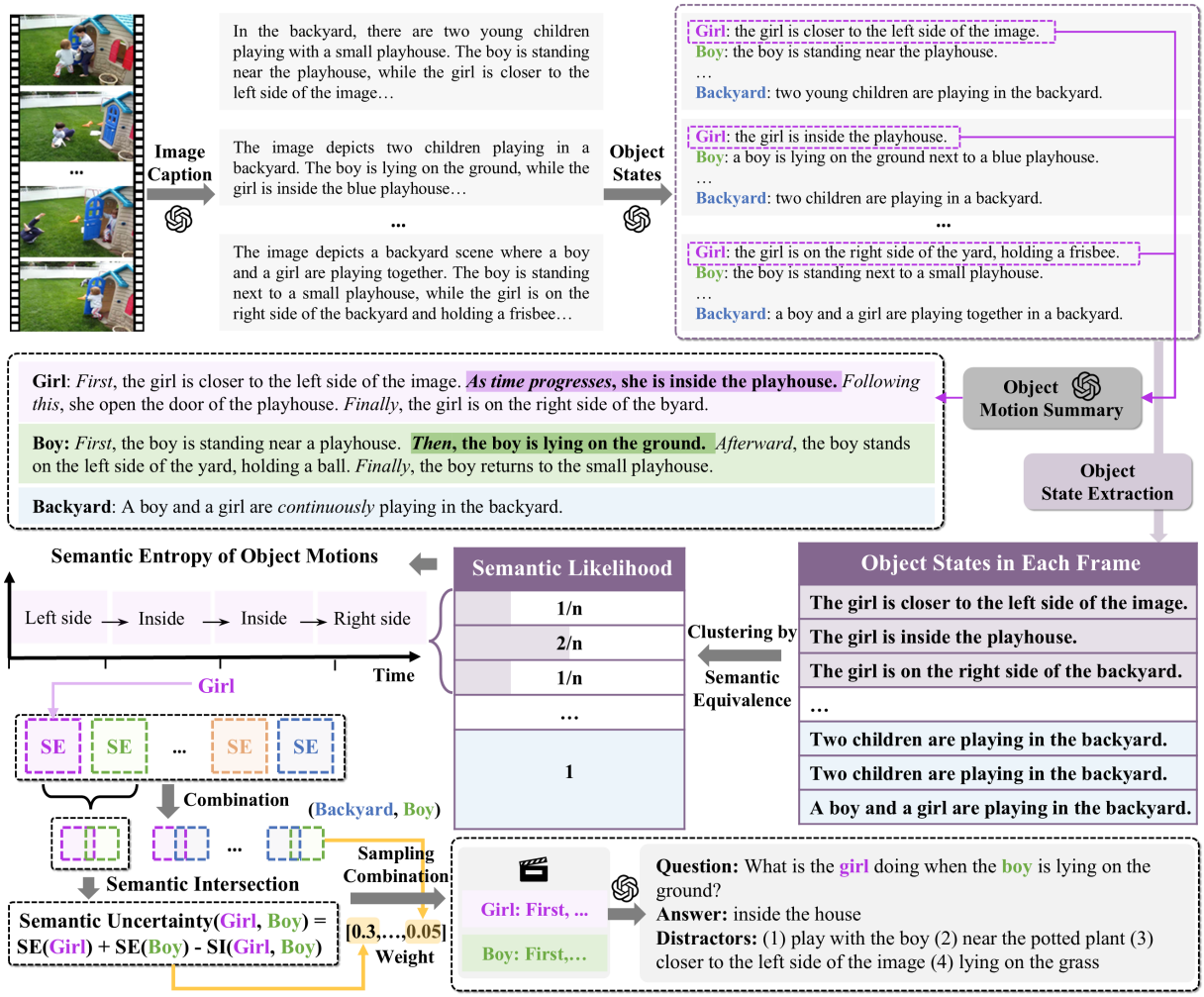

Figure 2: The model architecture.

explore the simultaneous generation of multiple QADs from a video, incorporating considerations of both consistency and diversity.

3 Method

3.1 Preliminary

Suppose we are going to generate n groups of QADs for the video V containing m objects, each of which exhibits motions and interacts with other objects in V. The object set is given as $O = \{o_1, ..., o_i, ...o_m\}$, and the temporal object motion set is $M = \{m_1, ..., m_i, ...m_m\}$, where each object o_i is associated with its temporal object motion m_i. We assume that each QAD_i is generated based on an object subset $O_i \subseteq O$ and a motion subset $M_i \subseteq M$, such as the generated QADs in Figure 2 which comprises the objects "girl" and "boy" and their motions. Then, we can define the generation of QAD_i as follows:

$$QAD_i = \mathcal{F}(V, O_i, M_i), \quad (1)$$

where $\mathcal{F}$ denotes the pre-trained vision-language model. The optimization objective is to maximize the total semantic information provided by n QADs for V as

$$\max \sum_{i=1}^{n} \mathcal{G}(QAD_i), \quad (2)$$

where $\mathcal{G}$ denotes the semantic information evaluation function.

3.2 Temporal Object Motions

Temporal object motions describe the varying states of objects throughout a video, playing a crucial role in determining objects' significance and semantic expressions in generating QADs. In this subsection, we will delineate the method of extracting temporal motions for each object.

For the video V, we first perform sparse sampling to obtain its T image frames. We then feed each image frame into a pre-trained large vision-language model to obtain image captions. Next,

we use a pre-trained large language model to extract objects from the image caption of each frame, ultimately forming the existing objects present in V. Later, we sequentially extract objects' descriptions from image captions across different frames, capturing the objects' changing states over time. Finally, we combine these states and employ a pre-trained large language model to generate the temporal motions for each object.

Formally, the temporal object motion for the object o_i can be defined as

$$tom_i = s\Big([I_1(o_i), ..., I_t(o_i), ..., I_T(o_i)]\Big), \quad (3)$$

where $I_t(o_i)$ denotes the state of o_i in the image frame I_t, and s denotes the summary of all object states. We use a large language model to generate s in chronological order using words that can express the temporal sequence, such as "*First*", "*Then*", and "*Finally*".

Examples of generating temporal object motions can be found in Appendix A. Although temporal object motions can be directly generated using LVLMs, our method achieves higher learning performance (Section 4.5).

3.3 Semantic Uncertainty

In this section, we propose a novel method based on semantic uncertainty to determine the semantic information of QADs. Intuitively, the QADs that contain rich semantic information are more enlightening and focus on the focal points in videos.

We assess objects' semantic uncertainty (Kuhn et al., 2023) to evaluate the amount of the semantic information given by QADs. That is, the objects that demonstrate substantial semantic variations throughout the video can be considered to have rich semantic content, offering great inspiration for generating QADs. Given QAD_i generated based on the object subset O_i, the total information of QAD_i can be defined as

$$\mathcal{G}(QAD_i) = SU(O_i), \quad (4)$$

where $SU(O_i)$ represents the semantic uncertainty of the object subset O_i. In this study, we leverage semantic entropy and semantic intersection to define $SU(O_i)$ as follows:

$$SU(O_i) = \sum_{o_i \in O_i} SE(o_i) - \sum_{o_i, o_j \in O_i} SI(o_i, o_j), \quad (5)$$

where $SE(o_i)$ represents the semantic entropy of the object o_i and $SI(o_i, o_j)$ denotes the semantic intersection of the two objects.

Quantifying the information content of individual motions or interactions is inherently complex. Drawing inspiration from (Farquhar et al., 2024), we assert that the motions or interactions of objects with higher semantic uncertainty are more significant and contain more information. Consequently, we propose to define semantic entropy to determine the semantic uncertainty of each individual object as follows:

$$SE(o_i) = -\sum_{\mathbf{C}_i} \Big(\Big[\sum_{m_i \in \mathbf{C}_i} p(m_i|o_i) \Big] \log \Big[\sum_{m_i \in \mathbf{C}_i} p(m_i|o_i) \Big] \Big), \quad (6)$$

where $\mathbf{C}_i$ represents a semantic equivalence class, and $p(m_i|o_i)$ is the conditional probability of the temporal motion m_i given o_i as the condition. In terms of the construction of semantic equivalence classes, we extract T temporal states $\{m_i^1, ..., m_i^t, ...m_i^T\}$ according to the distribution $p(m_i|o_i)$, where $m_i^t = I_t(o_i)$. The temporal states are clustered into $\mathcal{C}$ classes[1] based on their semantics, and each class captures the states that express the same meaning. Considering the temporal nature of object states, if there are other states present between two semantic equivalence states, we will not classify them into the same category. For example, "stand-stand-sit-stand" will be clustered as three classes rather than two classes. We ultimately obtain a temporal state set $\hat{m}_i = \{\hat{m}_i^1, ..., \hat{m}_i^{\mathcal{C}}\}$.

As to the semantic intersection between two objects, we identify all elements in $\hat{m}_i$ that are semantically equivalent as follows:

$$\hat{M} = \Big\{ (\hat{m}_i^l, \hat{m}_i^{l+1}) \Big| sim(\hat{m}_i^l, \hat{m}_j^k) \geq \theta \wedge sim(\hat{m}_i^{l+1}, \hat{m}_j^{k+1}) \geq \theta \Big\}, \quad (7)$$

where $l, k \in [1, \mathcal{C}-1]$ and $sim(\hat{m}_i^l, \hat{m}_j^k)$ is the similarity of two states. The set $\hat{M}$ comprises all of the elements for which the similarity between two consecutive elements in sets $\hat{m}_i$ and $\hat{m}_j$ exceeds a predefined threshold θ. Then, the semantic intersection of two objects can be defined as

$$SI(o_i, o_j) = -\sum_{\hat{m}_i, \hat{m}_j \in \hat{M}} p(\hat{m}_i|\hat{m}_j) \log p(\hat{m}_i|\hat{m}_j). \quad (8)$$

[1]The number of semantic equivalence classes is varying across different objects.

Finally, we can calculate the overall semantic uncertainty of all object combinations as

$$SU = [SU(O_1), ..., SU(O_{\mathcal{K}})], \tag{9}$$

where $\mathcal{K} = C_m^{|O_i|}$, representing the number of all possible object combinations extracted from O. We convert SU into sampling probabilities by normalizing SU as follows:

$$P(O) = \left[\, p(O_i) \,\right]_{i=1}^{\mathcal{K}},\ p(O_i) = \frac{SU(O_i)}{\sum_{j=1}^{\mathcal{K}} SU(O_j)}. \tag{10}$$

Given the video V containing n ground-truth QADs, we first sample n combinations of temporal object motions based on $P(O)$ to construct the candidate object set $O' = [O_1, ..., O_n]$. Then, we generate each QADs based on each $O_i \in O'$. Compared with directly selecting the top-n object combinations with the highest SU scores for QADs generation, our sample strategy ensures the nuanced interactions still have a chance to be captured by the model. When nuanced actions are captured in image captions, they become part of the object's motion record and remain eligible for selection through our mechanism. Considering the lack of a discernible correspondence between the generated QAD_i and the ground truth QAD_j^*, we draw inspiration from the evaluation methods in the recommendation field. That is, for each QAD_j^*, if its similarity to the most similar predicted result QAD_i exceeds a predefined threshold t, it can be considered as successfully predicted, formulated as

$$QAD^* = \begin{cases} QAD_{\arg\max S_{ij}} \ if\ S_{ij} > t \\ none \qquad\quad otherwise \end{cases}, \tag{11}$$

where S_{ij} denotes the cosine similarity of QAD_j^* and QAD_i. In the pilot study, we found that increasing the number of predicted QADs could significantly improve learning performance. Therefore, we will ultimately generate $r * n$ groups of QADs, where r is a predefined hyperparameter (its effects will be evaluated in Section 4.5).

4 Experiments

4.1 Datasets

NExT-QA. NExT-QA (Xiao et al., 2021) is a video question answering (VideoQA) benchmark that requires QA models to reason about causal and temporal actions and understand the rich object interactions in daily activities. We conduct the experiments on the test split of NExT-QA, which consists of 1,000 videos and 8,564 Multi-Choice QA pairs.

Perception Test. Perception Test (Pătrăucean et al., 2023) is a multimodal benchmark designed to evaluate the perception and reasoning skills of multimodal video models. We conducted the experiments on the train split of the Multiple-Choice Video QA subset. We removed QA pairs that included question types with very few occurrences. The final dataset consists of 1,401 videos and 5,460 Multi-Choice QA pairs.

4.2 Metrics

We evaluate the generated QADs from the perspectives of **consistency** and **diversity**. Given that consistency represents the concordance between the predicted results and the ground truth QADs, we utilize recall, precision, and the F1-score as the evaluation metrics. Recall can be calculated using $r = k/n$, we utilize BLEU (Papineni et al., 2002) and CIDEr (Vedantam et al., 2015) to evaluate the precision of the generated QADs. On the diversity side, we refer to (Wang et al., 2016) and select mBLEU-1 and mBLEU-4 as the evaluation metrics.

4.3 Baselines

We compare our model with the Large Language Models and Large Vision-Language Models. The model details can be found in Appendix B. Furthermore, we incorporated two image-based QAD frameworks as baselines for comparison. Specifically, in the case of VQADG (Ding et al., 2024a), a single image was randomly sampled from each video. For the ReBo (Ding et al., 2024b) framework, we adapted its recurrent generation component to OneVision. We use OneVision (Li et al., 2024a), LLaVA-NeXT (Liu et al., 2024), and mPLUG-Owl3 (Ye et al., 2024) to demonstrate the effectiveness and generalizability of our proposed methods[2]. The implementation details can be found in Appendix C. The source code of our model will be released once accepted.

4.4 Main Results

Overall Performance We evaluate the consistency and diversity of generated QADs on NExT-QA and Perception Test benchmarks in Table 1 and Table 2.

[2]We chose foundation models that can directly accept video input, so models like GPT-4o, Claude-3.5, and LLaMA-3.2 were not selected.

Model	Modality		Consistency(↑)				Diversity(↓)	
	Text	Visual	Recall	Precision		F1-Score	mBLEU-1	mBLEU-4
				BLEU-4	CIDEr			
GLM-4	✓	–	52.09	4.08	8.04	13.93	27.15	24.17
Qwen-2.5	✓	–	66.03	5.01	11.53	19.63	21.92	17.36
Qwen-2	✓	–	66.19	4.96	9.68	16.89	20.66	15.74
Llama-3	✓	–	66.63	4.87	9.48	16.60	20.57	15.80
ChatGPT	✓	–	68.36	5.10	9.71	17.00	19.54	14.46
Llama-3.1	✓	–	68.81	4.78	9.24	16.29	19.47	14.56
BLIP-3	✓	✓	8.98	5.51	13.36	10.74	58.44	57.45
Video-LLaVA	✓	✓	30.95	5.83	16.20	21.27	32.18	30.56
Qwen2-VL	✓	✓	33.15	4.82	12.42	18.07	32.21	31.74
Interleave	✓	✓	54.11	5.43	16.95	25.81	28.66	26.54
Qwen-VL	✓	✓	57.63	5.39	16.64	25.82	23.36	19.51
VQADG†	✓	✓	30.56	7.78	**50.13**	37.97	34.44	31.94
ReBo†	✓	✓	36.25	6.51	19.37	25.25	26.31	24.46
LLaVA-NeXT	✓	✓	57.65	6.57	20.08	29.79	25.33	23.35
+ *MaxSem*	✓	✓	**81.59**	**9.05**	**29.84**	**43.70**	**14.72**	**10.68**
mPLUG-Owl3	✓	✓	60.40	5.40	13.94	22.65	25.19	22.90
+ *MaxSem*	✓	✓	**83.93**	**7.32**	**18.96**	**30.93**	**15.45**	**11.17**
OneVision	✓	✓	65.48	6.40	19.60	30.17	24.43	22.52
+ *MaxSem*	✓	✓	**86.71**	**9.35**	**30.16**	**44.75**	**15.01**	**11.06**

Table 1: Performance evaluation on the NExT-QA dataset. We select CIDEr to calculate the F1-Score. "†" denotes our re-implementation.

Model	R	P	F_1	mB_4(↓)
GLM-4	22.64	23.86	23.23	21.11
Llama-3	26.12	20.95	23.25	14.60
ChatGPT	30.53	24.41	27.13	15.45
Qwen-2.5	37.80	30.81	33.95	12.58
BLIP-3	38.30	81.74	52.16	9.03
mPLUG-Owl3	39.92	83.32	53.98	6.28
Interleave	44.04	93.26	59.83	7.38
LLaVA-NeXT	32.69	36.59	34.53	13.25
+ *MaxSem*	**56.00**	**79.96**	**65.87**	**5.24**
Onevision	39.79	93.83	55.88	7.46
+ *MaxSem*	**54.42**	**101.72**	**70.91**	**5.50**

Table 2: Performance evaluation on the Perception Test dataset. R stands for recall, P stands for precision.

We provide video captions to instruct LLMs to generate QADs, and image frames or videos for LVLMs based on the model requirements. The detailed prompts used for LLMs and LVLMs are provided in Appendix A. There are several notable observations as follows:

- Our proposed method demonstrably outperforms baseline methods across all evaluated metrics. Notably, it achieves a maximum 23.97% recall improvement and a maximum 14.58% F1-Score improvement on the NeXT-QA benchmark. Similar performance gains are observed on the Perception Test benchmark. Substantial findings indicate that integrating temporal object motions significantly enhances the informational content of generated QADs, providing knowledge of both individual object movement and interaction between objects. Consequently, the incorporation of temporal object motions into the generation process yields QADs that more closely resemble those in the original benchmarks. Furthermore, the generation of multiple QADs incorporates a variety of combinations of temporal object motions, resulting in enhanced diversity for each video.
- Comparative analysis on the NExT-QA benchmark reveals that LLMs achieve higher recall and diversity scores, whereas LVLMs exhibit superior precision and F1 scores. On the Perception Test benchmark, LVLMs significantly outperform LLMs across all measured metrics. The inadequate representation of video information by captions is a likely contributing factor to the poor performance of LLMs.

Model	Question			Answer			Distractor		
	BLEU-4	CIDEr	$mB_1(\downarrow)$	BLEU-4	CIDEr	$mB_1(\downarrow)$	BLEU-4	CIDEr	$mB_1(\downarrow)$
GLM-4	8.19	30.42	27.85	0.39	13.45	25.86	0.80	5.39	26.06
Qwen-2.5	9.90	32.79	23.19	0.41	12.24	20.18	1.23	7.92	20.03
Llama-3	10.08	38.91	22.62	0.50	13.53	19.49	0.71	4.45	18.80
Llama-3.1	10.30	37.75	21.00	0.65	15.54	17.72	0.79	4.73	17.90
ChatGPT	10.35	36.51	21.03	0.42	13.21	17.78	0.74	5.68	17.60
Qwen-2	10.74	37.04	22.44	0.26	12.22	19.79	0.93	5.72	18.40
Qwen-VL	8.39	38.01	24.69	2.13	28.48	22.62	1.61	9.05	22.02
Qwen2-VL	8.53	43.46	32.30	0.89	16.54	31.28	0.85	4.56	32.26
BLIP-3	8.91	25.89	58.75	0.02	8.38	59.24	1.08	8.79	57.80
Interleave	9.28	39.15	28.74	1.67	21.75	27.26	1.38	9.11	28.30
Video-LLaVA	10.46	37.83	32.03	1.73	24.27	31.83	1.61	7.57	31.86
OneVision	10.47	45.65	24.44	1.39	24.96	23.53	1.70	9.22	24.16
+ *MaxSem*	**16.77**	**77.66**	**15.32**	**1.60**	**25.94**	**12.78**	**1.88**	**11.83**	**14.14**
mPLUG-Owl3	10.94	43.14	24.68	0.72	18.50	24.77	1.21	7.37	24.87
+ *MaxSem*	**16.48**	**72.11**	**15.83**	**0.73**	**19.74**	**12.87**	**1.29**	**8.29**	**14.44**
LLaVA-NeXT	11.22	40.78	25.48	1.49	19.88	24.07	1.73	10.56	24.95
+ *MaxSem*	**15.68**	**68.70**	**14.78**	**1.98**	**25.51**	**12.38**	**2.08**	**13.08**	**13.91**

Table 3: Separate evaluation of question, answer, and distractor on the NExT-QA dataset.

Separate Evaluation Table 3 presents the comparison of the automatic evaluation results of questions, answers, and distractors separately, which demonstrates that all the baselines fail to compete with our methods on both consistency and diversity. Specifically, we observe that improvements in answers and distractors were less pronounced than those in the questions. Therefore, we conducted a further evaluation of our generated QADs using the VQA task in Figure 3. We treat the generated questions as questions and input the generated answers and distractors as candidates into LVLMs. Subsequently, we calculated the accuracy of the models in correctly answering the generated questions. Given the inherent uncertainty in verifying the correctness of the generated answers, the following measures were implemented: (1) A "None of the above" option was included; (2) Multiple LLMs were employed for independent evaluation. Table 3 and Figure 3 indicate that our proposed method generates not only consistent and diverse questions, but also QADs with higher accuracy.

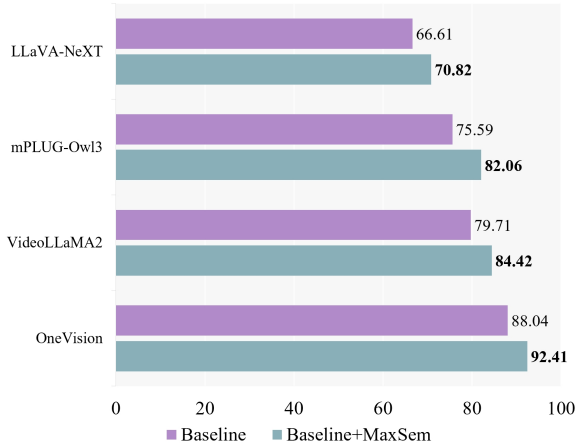

Figure 3: VQA task performance using different models on the dataset generated by Onevision.

4.5 Ablation Study

Object Count and Ratio We perform ablation studies to investigate the effects of the object count and ratio r, as presented in Figure 4. Our experimental findings demonstrate that the ratio parameter exerts a more substantial influence on the results than the object count parameter. This is a consequence of the variable number of objects present in the QADs within the real-world benchmarks. An object count of 2 and a ratio of 3 were ultimately selected.

Different Temporal Object Motion Recall rates were evaluated under four object motion conditions: (1) backbone model (no object motion); (2) whole object motion (all object motions included); (3) a simplified version (temporal object motions are generated directly from raw videos, and object sampling is performed randomly); (4) our proposed MaxSem model. As shown in Figure 5, the proposed MaxSem model yields the highest recall rate.

While employing information extracted from

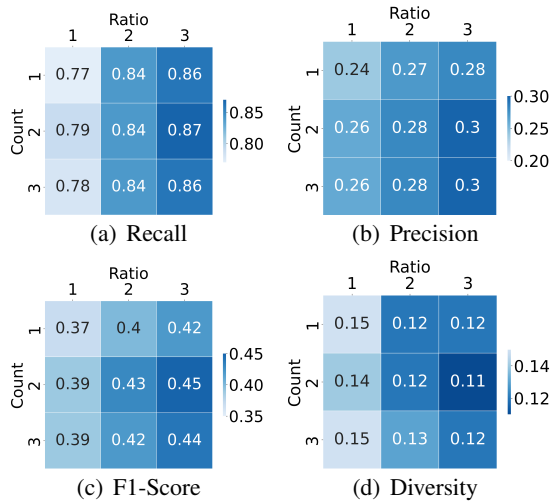

Figure 4: The evaluation results using different combinations of hyperparameters.

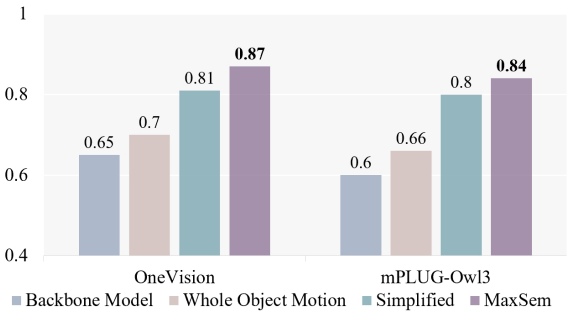

Figure 5: Results with different types of object motion.

multiple images and tracking the movements of multiple objects can aid in mitigating error propagation, the multi-step approach may nonetheless influence the final results. We provide the time cost of each step and the proposed method in Appendix D. Although our multi-step method will cost more time. We have developed two strategies to reduce computational costs: (1) reducing both the object count and ratio, as illustrated in Figure 4, and (2) leveraging video captions instead of image captions, as shown in the simplified example in Figure 5. We also evaluate how well the generated QADs reflect the temporal motions in Appendix E.

4.6 Human Evaluation

In addition, three annotators were engaged to assess 300 QADs generated by different methods. We adopt a 5 point scale for three metrics to evaluate the quality of generated QADs including: (1) Q, A, D_1: Overall quality of questions, answers, and distractors; (2) C: Consistency estimates whether the generated QADs are semantically relevant to the given ground truth QADs; (3) D_2: Diversity measures the degree of semantic divergence among the several QADs generated for a given video. Table 4 summarizes the human evaluation results of generated QADs. Our method significantly outperforms all baselines across all metrics, demonstrating that the generated QADs exhibit high quality, strong consistency with the benchmark dataset, and high diversity within the set of QADs generated for each video.

Model	Q	A	D_1	C	D_2
ChatGPT	3.55	3.62	3.4	3.39	3.68
Qwen-VL	3.78	3.59	3.51	3.21	3.45
OneVision	3.95	3.76	3.48	3.65	3.54
MaxSem	**4.12**	**3.89**	**3.64**	**4.05**	**4.2**

Table 4: Human evaluation of the generated QADs.

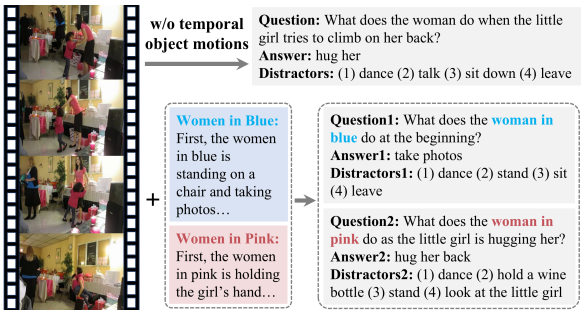

Figure 6: Case Study.

4.7 Case Study

To qualitatively evaluate the proposed method, we visualize an example from the NExT-QA benchmark in Figure 6. It shows from the figure that our model is capable of generating distinct QADs that specifically focus on *"women in blue"* and *"women in pink"*, respectively. In contrast, when the temporal object motions are removed, the model is only able to generate QADs that concentrate on segments of information from the video.

5 Conclusion

This paper presents MaxSem, a novel method that leverages temporal object motions and maximum semantic uncertainty to generate multiple QADs from a given video. Specifically, we use temporal object motions to describe the object movement over time and the intersections between different objects. Meanwhile, we introduce a selection module to select different temporal object motions to guide the generation. We conduct experiments on the NExT-QA and Perception Test benchmarks to

demonstrate a significant improvement of our proposed methods for different models.

6 Limitations

Our focus in this study is devoted on generating diverse QADs jointly for videos. The inherent challenge of this task arises from the necessity of comprehending both the temporal motions and interactions of objects within the video, as well as the generation and evaluation of QADs. We notice that there is still large room for progress. For example, how to represent the video information in a structured format, and how to efficiently process video data remain unaddressed and will be tackled in our future study. This study primarily focuses on general domains with abundant resources. In resource-constrained settings, transfer learning or supervised fine-tuning of pre-trained models may be necessary.

7 Acknowledgements

This work was supported in part by the National Natural Science Foundations of China (Nos. 62306156, 62106091), in part by the Undergraduate Education and Teaching Reform Projects, Nankai University (NKJG2025047), and in part by the Shandong Provincial Natural Science Foundation (No. ZR2021MF054).

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

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

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Wenjian Ding, Yao Zhang, Jun Wang, Adam Jatowt, and Zhenglu Yang. 2024a. Can we learn question, answer, and distractors all from an image? a new task for multiple-choice visual question answering. In *Proceedings of LREC-COLING*, pages 2852–2863.

Wenjian Ding, Yao Zhang, Jun Wang, Adam Jatowt, and Zhenglu Yang. 2024b. Exploring union and intersection of visual regions for generating questions, answers, and distractors. In *Proceedings of EMNLP*, pages 1479–1489.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Gemini Team Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of CVPR*, pages 6904–6913.

Zhaoyu Guo, Zhou Zhao, Weike Jin, Zhicheng Wei, Min Yang, Nannan Wang, and Nicholas Jing Yuan. 2020. Multi-turn video question generation via reinforced multi-choice attention network. *IEEE TCSVT*, 31(5):1697–1710.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of CVPR*, pages 2758–2766.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. 2020. Hierarchical conditional relation networks for video question answering. In *Proceedings of CVPR*, pages 9972–9981.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of EMNLP*, pages 1369–1379.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *Preprint*, arXiv:2407.07895.

Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2022. Invariant grounding for video question answering. In *Proceedings of CVPR*, pages 2928–2937.

Yicong Li, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-Seng Chua. 2023. Discovering spatio-temporal rationales for video question answering. In *Proceedings of ICCV*, pages 13869–13878.

Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. 2016. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of CVPR*, pages 4641–4650.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Jiaying Lu, Xin Ye, Yi Ren, and Yezhou Yang. 2022. Good, better, best: Textual distractors generation for multiple-choice visual question answering via reinforcement learning. In *Proceedings of CVPR*, pages 4921–4930.

Haohao Luo, Yang Deng, Ying Shen, See Kiong Ng, and Tat-Seng Chua. 2024. Chain-of-exemplar: Enhancing distractor generation for multimodal educational question generation. In *Proceedings of ACL*, pages 7978–7993.

OpenAI. 2024. Gpt-4o system card.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Proceedings of NeuIPS*, pages 27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, et al. 2023. Perception test: A diagnostic benchmark for multimodal video models. In *Proceedings of NeurIPS*.

Hung-Ting Su, Chen-Hsi Chang, Po-Wei Shen, Yu-Siang Wang, Ya-Liang Chang, Yu-Cheng Chang, Pu-Jen Cheng, and Winston H Hsu. 2021. End-to-end video question-answer generation with generator-pretester network. *IEEE TCSVT*, 31(11):4497–4507.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of CVPR*, pages 4566–4575.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Chen, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yu-Siang Wang, Hung-Ting Su, Chen-Hsi Chang, Zhe-Yu Liu, and Winston H Hsu. 2020. Video question generation via semantic rich cross-modal self-attention networks learning. In *Proceedings of ICASSP*, pages 2423–2427. IEEE.

Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. 2022. Language models with image descriptors are strong few-shot video-language learners. In *Proceedings of NeurIPS*, volume 35, pages 8483–8497.

Zhuhao Wang, Fei Wu, Weiming Lu, Jun Xiao, Xi Li, Zitong Zhang, and Yueting Zhuang. 2016. Diverse image captioning via grouptalk. In *Proceedings of IJCAI*, pages 2957–2964.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of CVPR*, pages 9777–9786.

Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. 2022a. Video as conditional graph hierarchy for multi-granular question answering. In *Proceedings of AAAI*, volume 36, pages 2804–2812.

Shaoning Xiao, Long Chen, Kaifeng Gao, Zhao Wang, Yi Yang, Zhimeng Zhang, and Jun Xiao. 2022b. Rethinking multi-modal alignment in multi-choice videoqa from feature and sample perspectives. In *Proceedings of EMNLP*, pages 8188–8198.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of ACM MM*, pages 1645–1653.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, et al. 2024. Blip-3: A family of open large multimodal models. *Preprint*, arXiv:2408.08872.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2021. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of ICCV*, pages 1686–1697.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *Preprint*, arXiv:2408.04840.

Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. 2017. Leveraging video descriptions to learn video question answering. In *Proceedings of AAAI*, volume 31.

A Prompts Details

A.1 Prompts for Temporal Object Motion Generation

Table 5 and Table 6 present the generation process of temporal object motions. For each object that appears in the video, we first use the Large Language Model to generate the object status for all image frames, then we use the Large Language Model to summarize the temporal object motions based on the generated object status.

Object Status Generation Prompt Input
Image caption: In the image, a baby girl is lying down on a couch with her back facing the camera. She is wearing a pink dress and appears to be enjoying herself. There are two people in the scene, one sitting on the left side of the couch and the other sitting on the right side of the couch. They both seem to be engaging with the baby girl, possibly playing with her or interacting with her in some way. **Object:** girl Refer to the following example and generate the status for the object based on the Image caption. The status should describe the condition or the action of the object, and the generated status should appear in the image caption. If the object does not appear in the image caption, then response 'Not appeared'. Only response the status, do not repeat the image caption and object. **Example:** **Image Caption:** The image features a man working at a desk in an office space. He is sitting at a desk with two computer monitors, one of which is placed closer to him. There are several books scattered around the room, some of which can be seen closer to the man. There is also a bottle of water on the desk, which could be used to hydrate the man as he works. Overall, the workspace appears to be well-organized and well-furnished. **Objects:** man **Status:** The man is working at a desk.

Table 5: Prompts used for object status generation.

Temporal Object Motions Generation Prompt Input
Object: woman **Object status 1:** a woman is present. **Object status 2:** a woman is on the right side of the image. **Object status 3:** a woman is sitting on the couch. **Object status 4:** a woman is holding a baby. **Object status 5:** a woman is observing the baby. **Object status 6:** a woman is sitting on the couch. Refer to the following example and generate the temporal object motion for the object based on all object status. The object status describe the actions or status of the object in a temporal sequence. Temporal object motion is generated by summarizing the object status in order using words that express temporal sequence, such as 'First', 'Then', 'Next', 'Later', and 'Finally'. Only response the temporal object motion. **Example:** **Object:** man **Object status 1:** a man is standing in front of a closed door. **Object status 2:** a man is standing next to the door. **Object status 3:** a man is standing in front of a closed door. **Object status 4:** the man's hand is on the door handle. **Object status 5:** a man reached for the door handle. **Object status 6:** a man is standing in front of a opened door. **Object status 7:** a man is in a room. **Object status 8:** a man is in a room. **Temporal object motion:** First, a man is standing in front of a closed floor. Next, a man reaches for the handle. Then, a man is standing in front of an opened door. Finally, a man is in a room.

Table 6: Prompts used for temporal object motion generation.

A.2 Prompts for QADs Generation

Table 7 presents the prompts used by the Large Language Models. For the Large Vision Language Models, we directly use video instead of video caption.

QADs Generation Prompt Input

Video caption: The video depicts a person with a large backpack walking along a dirt trail. The camera follows the person as they continue walking, and the view changes to show the surrounding landscape. The person appears to be on a journey, possibly hiking or trekking, as they navigate the trail. The backpack suggests that they may be carrying supplies for an extended trip. The trail itself seems to be in a rural or wilderness area, with no visible signs of civilization. The person's movements are deliberate and steady, indicating a level of experience and familiarity with the activity. Overall, the video captures the serene and peaceful atmosphere of a solitary journey through nature.

Refer to the following example and based on the video description, generate one question starting with 'what', and generate the answer and four distractors. The generated question should be related to the video description, the generated answer should be found in the video description, the distractors should be misleading but different from the answer, and the distractors should be separated with numbers like (1) (2) (3) (4).

Example:
Question: What does the girl do as the boy is attempting to put on the backpack at the end?
Answer: help him put on
Distractors: (1) pink (2) stop the boy from snatching (3) look towards the boy (4) disappointed

Table 7: Prompts used for QADs generation.

B Model Details

Large Language Models: We use ChatGPT (Ouyang et al., 2022), GLM4 (GLM et al., 2024), Llama3 (Dubey et al., 2024), Llama3.1 (Dubey et al., 2024), Qwen2 (Yang et al., 2024), and Qwen2.5 (Team, 2024) as baseline large language models.

Large Vision-Language Models: We use BLIP-3 (Xue et al., 2024), Video-LLaVA (Lin et al., 2023), Interleave (Li et al., 2024b), Qwen-VL (Bai et al., 2023) and Qwen2-VL (Wang et al., 2024) as baseline large vision language models.

Table 8 presents the specific model names in our experiments.

Model Name	Model Details
ChatGPT	gpt-3.5-turbo-0125
GLM-4	GLM4-9B-Chat
Llama-3	Lama3-8B-Instruct
Llama-3.1	Llama-3.1-8B-Instruct
Qwen-2	Qwen2-7B-Instruct
Qwen-2.5	Qwen2.5-7B-Instruct
BLIP-3	xGen-MM-instruct-interleave
Interleave	Llava-Interleave-Qwen-7B
Qwen-VL	Qwen-VL-Chat
Qwen2-VL	Qwen2-VL-7B-Instruct
Video-LLaVA	Video-LLaVA-7B
OneVision	LLaVA-OneVision-Qwen2-7B-ov
LLaVA-NeXT	LaVA-NeXT-Video-7B
mPLUG-Owl3	mPLUG-Owl3-7B-240728

Table 8: Baseline model details.

C Implementation Details

For temporal object motion generation, we sample 16 frames for each video. We use InstructBLIP-Vicuna-13B (Dai et al., 2023) to generate image captions, and we use Llama-3-8B-Instruct (Dubey et al., 2024) to generate existing objects, object s, and temporal object motions. We use the text-embedding-3-small model from OpenAI API for information gain and similarity calculation. We use the threshold $t = 0.6$ during evaluation. For Large Language Models, we first employ Video-LLaMA2 to generate video descriptions, and then we use these descriptions, rather than the videos, to generate QADs. For Large Vision-Language Models that can accept video inputs, we directly use the video to generate QADs. For large vision-language models that accept multiple image inputs, we sample 8 frames from the video and then use these frames to generate QADs. We use the HuggingFace[3] transformers library implementation for the above LLMs and LVLMs.

[3]https://huggingface.co/

D Computational Cost

We have measured the average time required to generate QADs for each video using both the baseline and MaxSem, as detailed in Table 9.

Average Time (s)	Onevision	mPLUG-Owl3
Baseline	42.63	65.28
MaxSem	192.96	218.71
Simplified	62.97	87.92

Table 9: Computational cost of our model.

We also present the time cost and the percentage of each step on OneVision in Table 10.

Content	Time (s)	Percentage (%)
Image Caption	41.85	21.69
Candidate Object	2.24	1.16
Object Status	61.52	31.88
Object Summary	7.49	3.88
Weight Generation	34.68	17.97
QAD Generation	45.18	23.41

Table 10: Computational cost of each step.

E Temporal Motion Evaluation

In order to evaluate how well the generated QADs reflect the temporal object motions of the video. We conduct a further evaluation by employing Llama3 to calculate the ratio of object motions mentioned in the QADs generated by various models. For each QAD, we iteratively assess its relevance to every object motion that we have generated for the video, with a higher ratio indicating that the generated QADs can better reflect the temporal motions present in the video.

Model	Onevision	mPLUG-Owl3	LLaVA-NeXT	Average
Baseline	87.07	87.81	80.15	85.01
MaxSem	97.19	97.52	91.85	95.52

Table 11: Temporal motion evaluation.