Towards Fine-Grained Video Question Answering

Wei Dai, Alan Luo, Zane Durante, Debadutta Dash, Arnold Milstein, Kevin Schulman,
Ehsan Adeli, Li Fei-Fei
Stanford University
450 Jane Stanford Way
{dvd.ai, alanzluo, durante, ddash, amilstein, Kevin.schulman, eadeli}@stanford.edu,
feifeili@cs.stanford.edu

Abstract

In the rapidly evolving domain of video understanding, Video Question Answering (VideoQA) remains a focal point. However, existing datasets exhibit gaps in temporal and spatial granularity, which consequently limits the capabilities of existing VideoQA methods. This paper introduces the Multi-Object Multi-Actor Question Answering (MOMA-QA) dataset, which is designed to address these shortcomings by emphasizing temporal localization, spatial relationship reasoning, and entity-centric queries. With ground truth scene graphs and temporal interval annotations, MOMA-QA is ideal for developing models for fine-grained video understanding. Furthermore, we present a novel video-language model, SGVLM, which incorporates a scene graph predictor, an efficient frame retriever, and a pre-trained large language model for temporal localization and fine-grained relationship understanding. Evaluations on MOMA-QA and other public datasets demonstrate the superior performance of our model, setting new benchmarks for VideoQA.

1. Introduction

In the current era of abundant digital video content, video understanding has become a key focus in computer vision research, with significant implications in various fields such as entertainment [10, 33, 43], healthcare [13, 32, 36, 37], and surveillance [40, 45]. Among the numerous aspects of video comprehension, Video Question Answering (VideoQA) has garnered a significant amount of attention, since it requires models to answer questions regarding a specific video segment, which necessitates a thorough grasp of the scene, relationships, and temporal changes depicted in the video [11, 51, 60].

Grounding in video understanding—specifically, temporal and spatial grounding—plays a pivotal role in bridging the gap between low-level video features and high-level semantic interpretations [48]. Temporal grounding ensures that events or actions within videos are associated with specific time intervals [38], while spatial grounding offers localized regions within video frames that correspond to certain entities or objects [41]. A dataset incorporating both temporal and spatial dimensions can offer rich contextual cues and pave the way for more detailed and accurate video comprehension tasks.

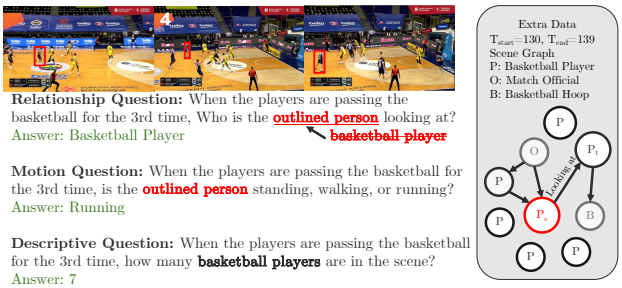


Figure 1. **Visualizations of Sample Questions from MOMA-QA.** We illustrate the three distinct types of questions in our dataset, each representing a different category for video question answering. All questions in our dataset are generated from a human-annotated spatio-temporal scene graph (shown on the right). The node of interest for the relationship and motion questions is colored red in the scene graph and outlined in the video.

While many datasets exist in the realm of video understanding, a deeper dive into their contents uncovers notable gaps in their representations of spatial relationships. For instance, in the ActivityNet-QA dataset [57], only 10% of its questions revolve around questions with spatial dimensions. This restricts the range and depth of inquiries a model can proficiently address. Another concern is the absence of fine-grained spatial relationship annotations. While TVQA+ [17] offers object-level details via bounding boxes, it fails to provide relationships between these objects. STAR [49] provides relationship annotations for its videos, but the automated nature of these annotations significantly restricts

their precision and applicability.

Temporal grounding, too, has its share of challenges in existing VideoQA datasets. Foundational datasets such as MSRVTT-QA [52] often neglect the importance of temporal localization. Approximately 33.4% of its questions can be distilled to a generic format: "What is [someone] doing" Such questions steer models predominantly toward video classification objectives, bypassing the need to anchor responses to specific moments or sequences within a video. Recent datasets like TVQA+ and TGIF-QA [11, 17] have shifted focus towards temporal reasoning within videos. However, they lack ground truth annotations for temporal localizations, thus there is no definitive means to ascertain whether a model has accurately localized the correct frames. Lastly, understanding and identifying the actions of individuals in crowded settings is challenging, and few datasets tackle this [26]. Addressing entity-specific queries in group situations is crucial for advanced video comprehension.

Given the gaps observed in current datasets, we introduce the Multi-Object Multi-Actor Question Answering (MOMA-QA) dataset. Stemming from the foundation of the *Multi-Object Multi-Actor (MOMA)* [26] dataset, MOMA-QA brings unique attributes designed to challenge and improve the current generation of video question-answering models. Firstly, as shown in Figure 1, every question within MOMA-QA requires temporal localization and is accompanied by ground truth temporal interval annotations to provide a means to validate models' temporal localization abilities. Secondly, 71.6% of the questions in the dataset require spatial relationship understanding, which MOMA-QA intensively assesses models on interpreting spatial connections among video entities. Each frame features ground truth scene graph annotations, laying a foundation for the evolution of more sophisticated spatially-aware models. Lastly, understanding the challenge of discerning specific individuals in crowded settings, we visually demarcate specific actors via frame-level bounding boxes on a subset of questions, thereby testing the model's proficiency in entity-specific reasoning.

As current datasets lack fine-grained annotations, existing VideoQA models struggle with nuanced understanding due to their linear approach of directly processing video frames and questions to produce answers. This limits their interpretability, a gap that becomes more apparent with the rise of visual language models in this domain. To address this issue, we introduce **SGVLM**, a video-language model with enhanced retrieval and relationship understanding abilities. Our model encapsulates three main features. First, the vision encoder has been restructured and augmented with a Motif-based scene graph generator [59]. The scene graph generator provides robust grounding for the spatial relationships depicted in videos and also provides an interpretable understanding of the model's decision-making pathway and elucidating how it arrives at its final predictions. Second, we devise an efficient frame retriever that identifies frames relevant to posed questions by leveraging both video and scene graph features, providing greater accuracy, especially for tasks on discerning relationships. Lastly, SGVLM harnesses the power of pre-trained large language models, empowering it to tackle intricate reasoning tasks.

In summary, our work has the following contributions: (1) We present the **MOMA-QA** dataset, a VideoQA dataset that emphasizes temporal localization, relationship reasoning through a vast array of questions, and frame-level entity-specific annotations to enhance video question-answering models. Each question is equipped with ground truth relationship and temporal annotation to facilitate the development of fine-grained VideoQA models. (2) We introduce **SGVLM**, a video-language model that features a restructured vision encoder with a Motif-based scene graph generator for spatial relationship grounding, an efficient frame retriever for selecting relevant frames, and the integration of pre-trained large language models for advanced reasoning capabilities.

2. Related Works

Video Question Answering Datasets. The quality of machine learning models is heavily influenced by the quality of their datasets. Foundational datasets like MovieQA [42], MSRVTT-QA, and MSVD-QA [52] have significantly advanced video question answering research [12, 20, 30, 54]. However, these datasets mainly include short clips with simple questions, limiting the development of models' in-depth video understanding. TGIF-QA [11] introduced a significant change by testing spatial-temporal reasoning in a large dataset of animated GIFs, leading to improvements in models' temporal reasoning abilities [5, 8]. Despite this, there remains a gap in spatial reasoning and the ability to handle crowded scenes with similar-looking actors.

Grounded VideoQA Models. Grounding in VideoQA tasks usually consists of two parts: spatial and temporal. With the advent of graph neural networks (GNN) [14], many works [9, 23, 29, 34] have integrated GNNs within their VideoQA framework for spatial grounding. Temporal grounding involves identifying salient frames related to the input question [3]. This technique gained increasing attention as LLM-based models became popular [1, 35, 44]. While LLM-based models bolster advanced reasoning abilities, their input lengths are strictly capped, making advance frame selection necessary [56]. This paper presents the first LLM-based VideoQA model that utilizes both temporal and spatial grounding features.

Category	Question Format	Answer Format	Example Question	Example Answer
Relationship	When $[C_i]$, what is $[V_s]$ $[E_i]$?	$[V_t]$	When the players are passing the basketball for the 3rd time, Who is the outlined person looking at?	Basketball Player
Motion	When $[C_i]$, is $[V_t]$ standing, walking or running?	[Att(V_t)]	When the players are passing the basketball for the 3rd time, is the outlined person standing, walking, or running?	Running
Description	When $[C_i]$, how many [Identity] are in the scene?	[Id. Count]	When the players are passing the basketball for the 3rd time, how many basketball players are in the scene?	7

Table 1. **General Structure of Generated Questions.** C_i denotes the description of a particular sub-activity. Additionally, V_s and V_t denotes the name of a source and a target node from the sub-activity. E_i represents the description of a relationship connecting V_s and V_t.

Dataset	Video Source	#Videos	#QA Pairs	Average Length (s)	Open Ended	Temporal Localization	Bounding Box Augmentation	Scene Graph Annotation
MSVD-QA [52]	MSVD	1,970	50,505	10	✓	✗	✗	✗
MSRVTT-QA [52]	MSRVTT	10,000	243,690	15	✓	✗	✗	✗
TGIF-QA [11]	TGIF	71,741	165,165	3	✓	✓	✗	✗
TVQA [16]	TV Show	21,793	152,545	76	✗	✓	✗	✗
ActivityNet-QA [57]	ActivityNet	5,800	58,000	180	✓	✓	✗	✗
Social-IQ [58]	YouTube	1,250	7,500	60	✗	✗	✗	✗
EgoSchema [27]	Ego4D	5,063	5,063	180	✗	✗	✗	✗
NExT-QA [51]	YFCC-100M	5,440	52,044	44	✓	✓	✗	✗
STAR [49]	Charades	23,013	60,206	11	✗	✓	✗	◯*
TVQA+ [17]	TV Show	4,198	29,383	61	✓	✓	✓	✗
MOMA-QA (Raw only)	MOMA	1,412	83,223	376	✓	✓	✗	✓
MOMA-QA (Aug.)	MOMA	27,586	300,791	144	✓	✓	✓	✓

Table 2. **Dataset Comparisons.** Our proposed dataset sets a new benchmark for open-ended, long VideoQA by providing extensive human annotations and a large number of QA pairs. Raw only: Includes raw videos only. Aug.: Includes both raw videos and box-augmented videos. * The graph annotations provided by the STAR dataset are automatically generated, while MOMA-QA's are human annotated.

3. MOMA-QA Dataset

In this section, we introduce the MOMA-QA dataset through three perspectives: source annotations, questions, and its feature of bounding box augmentations. We perform the same video-wise train/validation/test split as in the MOMA dataset. We then show the statistics of MOMA-QA and compare it with the current VideoQA datasets.

3.1. Annotations

The MOMA dataset contains human annotated *activity graphs* at the frame level. Specifically, each frame i is annotated with a graph $G_i = (V_i, E_i)$, where V_i contains a set of entities, along with their bounding boxes and attributes in the scene. E_i contains the relationships between the entities. In addition, consecutive frames are grouped into sub-activities. Sub-activity j has label $(T_{start,j}, T_{end,j}, C_j)$, where $T_{start,j}, T_{end,j}$ denotes the start and end of a particular activity, and C_j contains the description about the sub-activity. Such fine-grained human annotations make it ideal for question generation on relationships while also providing extra information for model grounding.

3.2. Bounding Box Augmentations

Analyzing crowded scenes poses challenges like question ambiguity. Consider the inquiry: "What is the basketball player looking at?" Posed within the context of a match involving ten participants, it becomes unclear to which player the question is directed. Nonetheless, entity-centric reasoning within such crowded scenes is critical across various domains. For instance, in a sports event, the analysis of a particular player's performance gathers considerable interest. To address this issue, we propose *bounding box augmentations*. This technique generates edited videos highlighting the focused entity using ground truth bounding box annotations from the MOMA dataset, as shown in Figure 1. This method effectively resolves the ambiguity, thereby facilitating entity-centric reasoning within dense scenes. Furthermore, we substitute specific entity designations (like "basketball player") with the more generic term "*outlined person*." This alteration serves to minimize the hints the question may provide regarding the answer, thus preventing the model from inferring the answer based solely on the phrasing of the question. For fair comparison to existing works, we do not supply any bounding box coordinates during QA.

3.3. Questions

In the MOMA-QA dataset, we offer three categories of questions: *relationship*, *motion*, and *description*. The standard template used to generate these questions is outlined in Table 1. After generation, each question undergoes a manual verification process to confirm its clarity and remove any ambiguity. Adjustments are made to the phrasing of questions to enhance their naturalness.

Figure 1 presents exemplar questions from the MOMA-QA dataset. Specifically, every question is accompanied by precise interval and scene graph ground truth annotations. We hope the inclusion of this information could drive the development of more intricate multimodal models.

3.4. Dataset Statistics

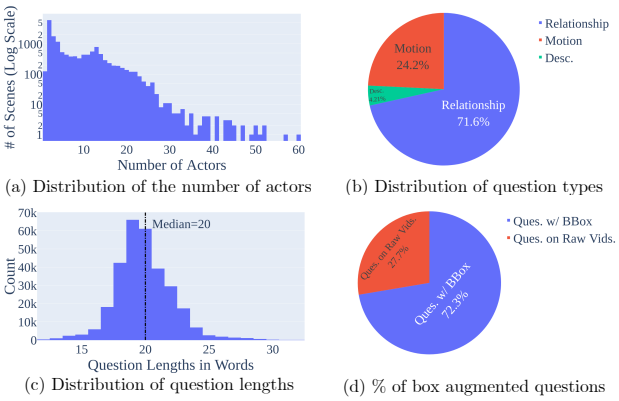
Table 2 presents a comparative analysis of the MOMA-QA dataset against other popular VideoQA datasets. The MOMA-QA dataset stands out with its extensive collection of 300,791 questions derived from 147 hours of original footage and an additional 956 hours of bounding-box augmented videos, making it one of the most comprehensive VideoQA datasets currently available. Furthermore, the average duration of 144 seconds per video makes this dataset well-suited for evaluating long-form temporal localization in models. In addition, MOMA-QA is one of the first datasets to provide human annotated temporal interval and scene graph data within the VideoQA domain.

Statistics from Figure 2 reveal that 71.6% of the questions in the dataset are centered on relationships, 24.2% pertain to motion, and the remaining 4.21% are descriptive. This distribution underscores the dataset's focus on relational understanding. Additionally, the dataset exhibits a balanced distribution of question lengths, with a median count of 20 words per question. It contains 4,045 scenes that contain over 10 actors each, and 72.3% of the questions have been enhanced with bounding box annotations, emphasizing the dataset's dedication to entity-specific queries.

4. Method

As shown in Figure 3, SGVLM fuses video frame features with language for advanced video understanding. Our model has five main components: frame encoder, scene graph predictor, frame localizer, Q-Former, and LLM. An illustration of each component is detailed below.

Frame Encoder. We utilize EVA-02 [6], a ViT based image encoder with 304M parameters, to generate patch image features $X \in \mathbb{R}^{L_{patch} \times d_v}$, object bounding box predictions $B = \{b_1, \ldots, b_n\}$, object class predictions $O = \{o_1, \ldots, o_n\}$, and image features used for scene graph generation. In particular, besides the coordinates of each box proposal prediction, the bounding box prediction B also contains a feature vector $\mathbf{f}_i$ and label probability $\mathbf{p}_i$ for each



(a) Distribution of the number of actors (b) Distribution of question types

(c) Distribution of question lengths (d) % of box augmented questions

Figure 2. **Statistics of MOMA-QA.** (a) The distribution of the number of actors. (b) The percentage of each question type in MOMA-QA. (c) The distribution of question lengths in MOMA-QA in words. (d) The percentage of box-augmented questions.

proposal $b_i \in B$. We utilize the pre-trained weight from the original work and fine-tune it on Visual Genome [15] and MOMA-QA.

Scene Graph Predictor. We design our scene graph predictor based on the Neural Motifs structure [59] to pre-train our frame encoder on Visual Genome [15]. We use biLSTM layers to encode object and edge contexts, which are then used to build relationship features. The features with top k probabilities $\mathbf{S} = \{\mathbf{s}_i, \ldots, \mathbf{s}_k\}$ are extracted and used in subsequent steps. A detailed explanation of the process is included in Suppl. A.

Frame Localizer. The frame localizer, based on the UniVTG [22] structure, employs a hybrid alignment and contrastive approach, leveraging frame and scene graph embeddings. During training, frames are labeled with binary f_i, where $f_i = 1$ signifies a foreground clip, and a saliency score $s_i \in [-1, 1]$, indicating relevance to the target question. The input question is converted into query tokens $\mathbf{Q} \in \mathbb{R}^{n \times d_t}$. For each frame, the frame embedding $\mathbf{X}$ and scene graph embedding $\mathbf{S}$ undergo a separate linear layer:

$$\mathbf{x}_i = \frac{1}{|\mathbf{X_i}|} \sum_{j=1}^{|\mathbf{X_i}|} \mathbf{X_i} \mathbf{W}_{xs}, \quad \mathbf{s}_i = \frac{1}{|\mathbf{S_i}|} \sum_{j=1}^{|\mathbf{S_i}|} \mathbf{S_i} \mathbf{W}_{ss},$$

where $\mathbf{W}_{xs}, \mathbf{W}_{ss}$ are learnable matrices. The squashed frame embedding and scene graph embeddings are then separately concatenated to form video frame embedding $\mathbf{X}_v = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and video scene graph embedding $\mathbf{S}_v = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ for video of length n. In the alignment route, each modality is appended with the positional embedding and type embeddings: $\mathbf{X}'_v = \mathbf{X}_v + \mathbf{E}^{pos}_X + \mathbf{E}^{type}_X$; $\mathbf{Q}'_v = \mathbf{Q}_v + \mathbf{E}^{pos}_Q + \mathbf{E}^{type}_Q$; $\mathbf{S}'_v = \mathbf{S}_v + \mathbf{E}^{pos}_S + \mathbf{E}^{type}_S$. Then, the embeddings concatenated into $\mathbf{Z}_0 = [\mathbf{X}'_v; \mathbf{S}'_v; \mathbf{Q}'_v]$. The concatenated representation $\mathbf{Z}_0$ is fed into a stack of k transformer encoders, where each encoder is composed of a

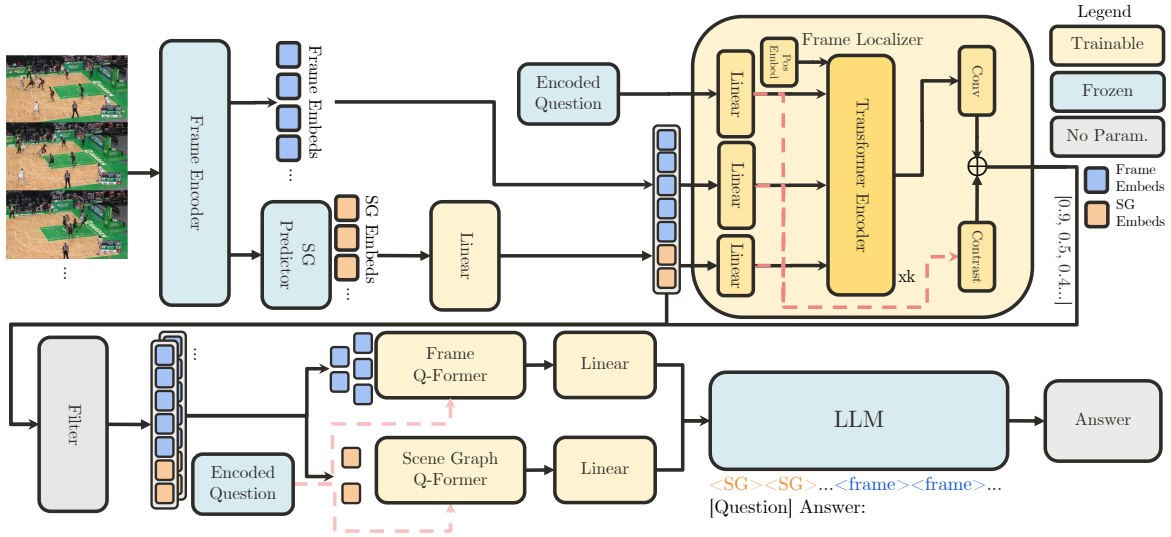


Figure 3. **Model Architecture of SGVLM.** The model employs a frame encoder to extract frame embeddings from the input video, which are subsequently used by a Scene Graph (SG) Predictor to generate scene graph embeddings. These embeddings are then concatenated with the frame features. The combination, along with question embeddings, is processed by a transformer encoder in the Frame Localizer to produce similarity scores for identifying relevant frames. Key frame features are then processed by Frame Q-Former and SG Q-Former to align with the language query and scene graph features. An LLM finally generates answers using a structured representation of scene graph and frame data, merged with the natural language question.

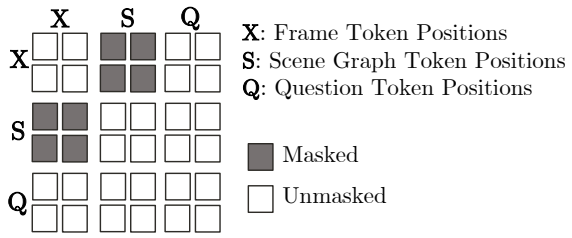


Figure 4. **Self-Attention Mask of the Transformer Encoder in Frame Localizer.** To separate frame and scene graph tokens, we mask out portions of the input with $-\infty$.

multi-head self-attention (MHSA) and a linear layer. At layer i with an MHSA with m heads, we have

$$\mathbf{k}_{i,m} = \text{softmax}\left(\frac{\mathbf{W}_Q^{i,m}\mathbf{Z}_{i-1}(\mathbf{W}_K^{i,m}\mathbf{Z}_{i-1})^T}{\sqrt{d_k^i}} + \mathbf{M}\right),$$
$$\mathbf{h}_{i,m} = \mathbf{k}_{i,m}\mathbf{W}_V^{i,m}\mathbf{Z}_{i-1},$$
$$\mathbf{Z}_i = (||_{m=1}^{M}\mathbf{h}_{i,m})\mathbf{W}_O^i,$$

where $\mathbf{W}_O^i, \mathbf{W}_Q^{i,m}, \mathbf{W}_K^{i,m}, \mathbf{W}_V^{i,m}$ are learnable parameters, and $\mathbf{M}$ is the attention mask as configured in Figure 4. Our architecture implements a specialized attention mask that restricts frame and scene graph tokens to interact exclusively with question tokens. This design choice is grounded in the empirical finding that scene graph and frame tokens exhibit inherent correlation. Without this masking, the localizer shows a propensity to assign high attention scores to the interplay between frame and scene graph tokens, often at the expense of question token relevance. By enforcing this attention mask, we ensure that the focus remains on integrating the question context effectively, as demonstrated by Suppl. D. In the end, we remove the question token part of $\mathbf{Z}_k$ and leave only the frame and scene graph tokens to obtain $\mathbf{Z}'_k$. The final score for the alignment route is then obtained by

$$\hat{\mathbf{f}} = \sigma(\text{Conv}(\mathbf{Z}'_k)), \tag{1}$$

where σ is a sigmoid activation, and Conv is a set of convolutional layers that outputs $\hat{\mathbf{f}} \in \mathbb{R}^n = \{\hat{f}_1, \ldots, \hat{f}_n\}$, where each value predicts whether the frame belongs to a foreground clip. The alignment route is then supervised by the cross entropy loss between the predicted label $\hat{\mathbf{f}}_a$ and the ground truth label f_a:

$$\mathcal{L}_a = \sum_{i=1}^{n} -\Big(f_i \log \hat{f}_i + (1-f_i)\log(1-\hat{f}_i)\Big), \tag{2}$$

where s_i is the ground truth relevance at frame i.

In the contrastive learning route, a one-layer attention layer is first used to project the question embedding $\mathbf{Q}' = \text{softmax}(\mathbf{W}_c\mathbf{Q})\mathbf{Q}$ where $\mathbf{W}_c$ is a learnable parameter. Then, the saliency score $\hat{s}_c$ is obtained through the sum of the pair-wise similarity score between the frame embedding $\mathbf{S}_v = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, scene graph embedding $\mathbf{X}_v = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, and question embedding $\mathbf{Q}'$:

$$\hat{s}_{c,i} = \frac{\mathbf{x}_i^T\mathbf{Q}'}{||\mathbf{x}_i||_2||\mathbf{Q}'||_2} + \frac{\mathbf{s}_i^T\mathbf{Q}'}{||\mathbf{s}_i||_2||\mathbf{Q}'||_2}. \tag{3}$$

This score is supervised through two losses: intra-video and inter-video contrastive learning loss. For intra-video contrastive learning loss, we randomly sample a positive clip at index p with $f_p = 1$ and $s_p > 0$, and negative samples $N = \{j | 1 \leq j < p, s_j < s_p\}$. Given the saliency prediction $\hat{s}_j, \hat{s}_j$, the intra-video loss is calculated as

$$\mathcal{L}_s^{\text{intra}} = -\log \frac{\exp(\hat{s}_p/\tau)}{\exp(\hat{s}_p/\tau) + \sum_{j \in N} \exp(\hat{s}_j/\tau)}, \quad (4)$$

where τ is a hyperparameter representing the temperature. The inter-video loss takes other videos $k \in N'$ within the batch as negative samples

$$\mathcal{L}_s^{\text{inter}} = -\log \frac{\exp(\hat{s}_p/\tau)}{\sum_{k \in B} \exp(\hat{s}_p^k/\tau)}. \quad (5)$$

The overall training objective is the weighted combination:

$$\mathcal{L} = \lambda_a \mathcal{L}_a + \lambda_{intra} \mathcal{L}_s^{\text{intra}} + \lambda_{inter} \mathcal{L}_s^{\text{inter}}, \quad (6)$$

where $\lambda_a, \lambda_{inter}, \lambda_{intra}$ are hyperparameters setting the weight for each loss. Finally, the relevance score r_i for frame i is the sum of both foreground prediction $\hat{f}_i$, and the saliency score $\hat{s}_i$:

$$\hat{r}_i = w_f \hat{f}_i + w_s \hat{s}_i, \quad (7)$$

where w_f, w_s are two learnable scalars representing the weight of each score. The frames are ranked based on $\hat{r}_i$, and only the top k frames are input into the Q-Formers in the next stage. For datasets with ground truth interval annotation (like MOMA-QA), the localizer is directly tuned on the ground truth labels. For datasets without ground truth labels, pseudo labels f_i', s_i' are generated to fine-tune the localizer. Specifically, for each frame i with answer prediction $y, \hat{y}$ and frame selection threshold r_θ,

$$f_i', s_i' = \begin{cases} 1, 1 & \text{if } (y = \hat{y} \wedge \hat{r}_i > r_\theta) \\ & \quad \vee (y \neq \hat{y} \wedge \hat{r}_i < r_\theta) \\ 0, -1 & \text{Otherwise} \end{cases}. \quad (8)$$

In other words, we encourage the localizer to make the same prediction if such prediction gives the correct answer while encouraging the model to make a different prediction when the selected frames fail to provide the correct answer.

Q-Formers and LLM. We implement the Q-Formers as designed in BLIP-2 [21]. In particular, since the scene graph and the frame embeddings have distinctively different embeddings, two Q-Formers are used, with one taking the frame embeddings and one taking the scene graph embeddings. A linear projection is then used to project the embedding into the LLM embedding space. Finally, the scene graph tokens, frame tokens and question tokens are concatenated and input into an LLM, and an LLM inference is performed to obtain the final answer.

5. Experiments

We evaluate our model against current state-of-the-art VideoQA models on MOMA-QA and two public datasets: NExT-QA and QVHighlights.

5.1. Dataset & Metrics

The MOMA-QA dataset is evaluated on two metrics: **Accuracy** and **WUPS@0.9**. As MOMA-QA's questions are open ended, with test dataset $\mathbf{Q}$, the accuracy of the prediction $\hat{q}$ with respect to ground truth q is given by:

$$acc = \frac{1}{|\mathbf{Q}|} \sum_{\mathbf{Q}} \frac{1}{|q|} \sum_{i=1}^{min(|\hat{q}|,|q|)} \mathbf{I}[\hat{q}_i = q_i]. \quad (9)$$

WUPS is a soft measurement of accuracy used in multiple recent VideoQA datasets [51, 57]. The calculation method is detailed in Suppl. B.

The models are also evaluated on two public datasets: **NExT-QA** and **QVHighlights**. NExT-QA [51] is a VideoQA dataset focusing on causal and temporal action reasoning with 5,440 videos and 52,044 multiple-choice questions grouped into three categories: temporal, causal, and descriptive. We report categorical and overall accuracy. QVHighlights [19] is a unified dataset for both moment retrieval and highlight detection. It contains 10,310 questions associated with 18,367 moments in 10,148 videos. We follow the evaluation metrics on the original paper and report R1, mAP on moment retrieval, and mAP and HIT@1 on highlight detection.

5.2. Experimental Setup

We evaluate the performance of current state-of-the-art models on the datasets in both zero-shot and fine-tuned contexts. The MOMA-QA results are compared against 4 popular models: InternVideo [47], mPLUG-2 [53], BLIP-2 [21], and SeViLa [56]. The details of the experimental setup and training process are included in Suppl. C.

6. Results & Discussions

In this section, we discuss the zero shot and fine-tuned VideoQA performance of various models on MOMA-QA and NeXT-QA. We also report the results of the retriever alone on QVHighlights. Finally, we conduct a qualitative analysis of the results reported by SGVLM.

6.1. Zero Shot Results

Table 3 shows the zero-shot metrics for the tested models. In our experiment, open vocabulary models surpass closed vocabulary ones in accuracy and WUPS, yet overall performances remain subpar. The best accuracy and WUPS@0.9 are 27.94% (SGVLM) and 0.6023 (SeViLa) respectively. However, even open vocabulary models show

Model	Description		Relationship		Action		Total	
	Accuracy	WUPS@0.9	Accuracy	WUPS@0.9	Accuracy	WUPS@0.9	Accuracy	WUPS@0.9
InternVideo [47]	0.00	0.0000	0.07	0.0018	0.00	0.0000	0.05	0.0013
mPLUG-2 [53]	0.00	0.5084	9.83	0.2891	0.00	0.0000	6.90	0.2222
BLIP-2 [21]	55.19	0.5519	12.30	0.1957	62.10	0.6210	26.45	0.3161
SeViLa [56]	53.22	**0.6027**	12.99	**0.5045**	64.66	0.8977	27.44	**0.6023**
SGVLM	**58.69**	0.5869	**13.03**	0.4663	**65.43**	**0.9174**	**27.94**	0.5828

Table 3. **Zero-Shot Performance Comparison of SGVLM with Baselines on MOMA-QA Dataset.** Our method outperforms, or performs on-par with existing methods in the zero-shot setting.

Model	Description		Relationship		Action		Total	
	Accuracy	WUPS@0.9	Accuracy	WUPS@0.9	Accuracy	WUPS@0.9	Accuracy	WUPS@0.9
InternVideo [47]	42.15	0.4215	36.77	0.3980	71.12	0.7112	45.91	0.4804
mPLUG-2 [53]	0.94	0.0094	47.00	0.7084	0.39	0.0056	33.12	0.4990
BLIP-2 [21]	62.90	0.6290	77.34	0.7864	72.55	0.9286	75.73	0.8278
Sevila [56]	63.60	0.6360	78.92	0.8218	74.60	0.9453	77.19	0.8442
SGVLM$_{NoLoc}$	**67.01**	**0.6701**	79.25	0.8216	74.71	0.9560	77.60	0.8482
SGVLM$_{NoSG}$	65.33	0.6533	78.73	0.8263	76.33	0.9763	77.55	0.8558
SGVLM	66.64	0.6664	**81.36**	**0.8435**	**77.06**	**0.9771**	**79.66**	**0.8688**

Table 4. **Fine-tuned Performance Comparison of SGVLM with Baselines on MOMA-QA Dataset.** SGVLM$_{NoLoc}$: An ablation of SGVLM where the frame localizer is removed and replaced with uniform frame sampling. SGVLM$_{NoSG}$: An ablation of SGVLM where the scene graph predictor is removed, and the model inferences solely on the frame embeddings.

Model	Causal	Temporal	Descriptive	Average
HGA [12]	46.8	52.1	59.3	50.4
All-in-One [46]	48.0	48.6	63.2	50.6
Just Ask [54]	49.6	51.4	63.1	52.3
MIST [7]	54.6	56.6	66.9	57.2
HiTeA [55]	62.4	58.3	75.6	63.1
InternVideo [47]	62.5	58.5	75.8	63.2
BLIP-2 [21]	72.9	68.1	81.2	72.6
SeViLA [56]	74.2	**69.4**	81.3	73.8
SGVLM	**75.2**	66.3	**83.4**	**74.3**

Table 5. **Comparison of SGVLM with SoTA on NExT-QA.** We achieve comparable or slightly superior performance to existing methods on the NExT-QA dataset. This is noteworthy considering NExT-QA lacks explicit relationship and scene-graph oriented questions, underscoring the versatility of our approach.

limited zero-shot performance, with a maximum accuracy of 13.03% and WUPS of 0.5045. These findings suggest a significant disparity between the MOMA-QA dataset and the datasets on which these models were originally trained, highlighting that current VideoQA models struggle with the intricate relational dynamics featured in MOMA-QA, regardless of their model structures. The distinctive characteristics of MOMA-QA therefore underscore its potential to introduce valuable diversity to the spectrum of VideoQA datasets available for advancing the field.

Model	Moment Retrieval					HD	
	R1		mAP			$\geq$ Very Good	
	@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
BeautyThumb [39]	–	–	–	–	–	14.36	20.88
DVSE [24]	–	–	–	–	–	18.75	21.79
MCN [2]	11.41	2.72	24.94	8.22	10.67	–	–
CAL [4]	25.49	11.54	23.40	7.65	9.89	–	–
CLIP [31]	16.88	5.19	18.11	7.0	7.67	31.30	61.04
XML [18]	41.83	30.35	44.63	31.73	32.14	34.49	55.25
XML+ [19]	46.69	33.46	47.89	34.67	34.90	35.38	55.06
MDETR [19]	52.89	33.02	54.82	29.40	30.73	35.69	55.60
UniVTG [22]	58.86	40.86	57.60	35.59	35.47	38.20	60.96
UMT [25]	56.23	41.18	53.83	37.01	36.12	38.18	59.99
QD-DETR [28]	62.40	44.98	**62.52**	39.88	**39.86**	38.94	**62.40**
SeViLA [56]	54.50	36.50	–	–	32.30	–	–
SGVLM	**63.36**	**46.30**	62.47	**42.00**	39.82	**39.17**	62.26

Table 6. **Moment Retrieval and Highlight Detection Results on QVHighlights Test Split.** We only include models not trained on additional video retrieval datasets (no extra training data). SGVLM (ours) and SeViLA are the only two VideoQA models.

6.2. Fine-tuned VideoQA Results

MOMA-QA. Table 4 presents a performance comparison of SGVLM with several baselines on the MOMA-QA Dataset. SGVLM outperforms the baseline methods across all metrics. Notably, in the Description and Relationship categories, SGVLM achieves accuracy scores of 66.64% and 81.36%, with corresponding WUPS@0.9 scores of 0.6664 and 0.8435, respectively. This represents a signif-

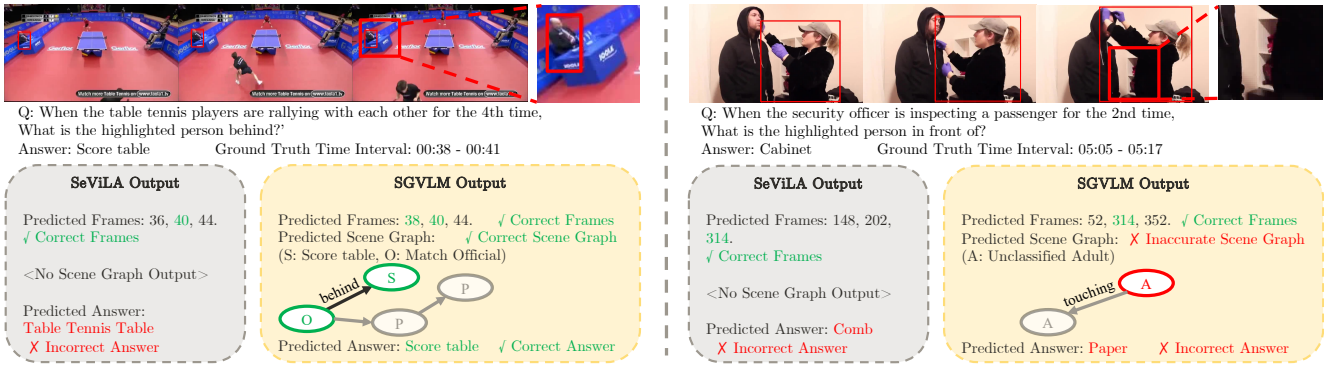


Figure 5. **Visualization Results of SGVLM with Previous SoTA (SeViLA) on MOMA-QA.** Left: An example where SGVLM makes the correct prediction while SeViLA fails. Right: An example where both our model and SeViLA produce incorrect answers. We magnify the part from the frame that is relevant to the question for better readability.

icant improvement of up to 3.04% over the SeViLA model. Overall, SGVLM demonstrates the highest total accuracy at 79.66% and a WUPS@0.9 score of 0.8688, suggesting robust video understanding performance across tasks.

NeXT-QA. As shown in Table 5, SGVLM outperforms existing models on NeXT-QA. In the causal and descriptive questions, SGVLM sets new records with accuracies of 75.2% and 83.4%, respectively, exceeding the previous SoTA by up to 2.1%. On average, SGVLM achieves an accuracy of 74.3%, demonstrating its superior performance across different video understanding challenges.

Ablations. An ablation study of each component is shown in Table 4, where we test how the model performs without the localizer and scene graph component. The ablations indicate that both parts contribute significantly to the model's performance. The SGVLM without the localizer (SGVLM$_{NoLoc}$) and without the scene graph predictor (SGVLM$_{NoSG}$) show reduced accuracy and WUPS@0.9 scores across all question categories (except *description*) compared to the complete SGVLM model. Specifically, SGVLM$_{NoLoc}$ shows a slight decrease across most categories, while SGVLM$_{NoSG}$ shows a more pronounced decrease in the Relationship category, suggesting the scene graph predictor helps the model the most in the relationship category. These results underscore the importance of both frame localization and scene graph predictions in driving the model's superior performance.

6.3. Frame Localization Results

Our SGVLM model exhibits strong performance in Moment Retrieval and Highlight Detection tasks as shown in Table 6. SGVLM's capabilities are particularly evident in moment retrieval, where it tops the charts, exceeding the closest competitor by as much as 2.12% in R1@0.5, R1@0.7, and mAP@0.7. In highlight detection, our model ranks second in mAP and leads in HIT@1, showcasing its precision in identifying video segments of interest. Notably, it outperforms the previously established state-of-the-art VideoQA model, SeViLA, by up to 9.8%. These findings confirm SGVLM's effectiveness in accurately locating relevant video moments, highlighting its potential for real-world video analysis applications.

6.4. Qualitative Analysis

In our qualitative analysis, we compare SGVLM's output with the prior SoTA, SeViLA. Figure 5 (left) illustrates the task of identifying the object in front of the highlighted match official. SGVLM not only accurately selects relevant frames but also successfully constructs a scene graph, correctly recognizing the *match official* as positioned *behind* the *score table*. In contrast, SeViLA, while identifying salient frames correctly, misinterprets the object as a *table tennis table*. In this scenario, the use of scene graphs in SGVLM evidently contributes to its enhanced reasoning capabilities, which affirm the utility of structured semantic representations in complex VideoQA tasks.

In the right portion of Figure 5, both models were assessed for their ability to correctly identify the object behind the highlighted person during a security inspection. Despite neither model successfully identifying the 'Cabinet' as the correct answer, SGVLM provides enhanced interpretability through its scene graph representation. The *cabinet* is missing from SGVLM's scene graph, which suggests a limitation in the vision encoder's capability to recognize this occluded object. This interpretative result directs attention to potential enhancements in the vision encoding component of the model, indicating a clear pathway for future improvements in the model's overall ability.

7. Conclusion

In this work, we introduce MOMA-QA, a VideoQA dataset that we hope will serve as a useful tool for advancing the fine-grained capabilities of VideoQA models by providing comprehensive frame-level annotations

for spatio-temporally grounded QA. Towards this end, we introduce a novel video-language model, referred to as SGVLM. Our model uniquely leverages MOMA-QA's scene graph annotations for precise spatial relationship understanding and temporal localization annotations for effective frame selection. By integrating fine-grained video understanding with pre-trained large language models, we achieve a new state-of-the-art for VideoQA.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pages 5803–5812, 2017. 7

[3] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2917–2927, 2022. 2

[4] Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell. Temporal localization of moments in video collections with natural language. *arXiv preprint arXiv:1907.12763*, 2019. 7

[5] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *CVPR*, pages 1999–2007. Computer Vision Foundation / IEEE, 2019. 2

[6] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023. 4

[7] Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14773–14783, 2023. 7

[8] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *CVPR*, pages 6576–6585. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[9] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. Location-aware graph convolutional networks for video question answering. In *AAAI*, pages 11021–11028. AAAI Press, 2020. 2

[10] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 709–727. Springer, 2020. 1

[11] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017. 1, 2, 3

[12] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *AAAI*, pages 11109–11116. AAAI Press, 2020. 2, 7

[13] Jalaluddin Khan, Jian Ping Li, Bilal Ahamad, Shadma Parveen, Amin Ul Haq, Ghufran Ahmad Khan, and Arun Kumar Sangaiah. Smsh: Secure surveillance mechanism on smart healthcare iot system with probabilistic image encryption. *IEEE Access*, 8:15747–15767, 2020. 1

[14] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR (Poster)*. OpenReview.net, 2017. 2

[15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 4

[16] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *EMNLP*, 2018. 3

[17] Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. TVQA+: spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8211–8225. Association for Computational Linguistics, 2020. 1, 2, 3

[18] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*, pages 447–463, 2020. 7

[19] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In *NeurIPS*, pages 11846–11858, 2021. 6, 7

[20] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, pages 7331–7341. Computer Vision Foundation / IEEE, 2021. 2

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 19730–19742. PMLR, 2023. 6, 7, 1

[22] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *CVPR*, pages 2794–2804, 2023. 4, 7

[23] Fei Liu, Jing Liu, Weining Wang, and Hanqing Lu. HAIR: hierarchical visual-semantic relational reasoning for video question answering. In *ICCV*, pages 1678–1687. IEEE, 2021. 2

[24] Wu Liu, Tao Mei, Yongdong Zhang, Cherry Che, and Jiebo Luo. Multi-task deep visual-semantic embedding for video thumbnail selection. In *CVPR*, pages 3707–3715, 2015. 7

[25] Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *CVPR*, pages 3042–3051, 2022. 7

[26] Zelun Luo, Zane Durante, Linden Li, Wanze Xie, Ruochen Liu, Emily Jin, Zhuoyi Huang, Lun Yu Li, Jiajun Wu, Juan Carlos Niebles, et al. Moma-lrg: Language-refined graphs for multi-object multi-actor activity parsing. *Advances in Neural Information Processing Systems*, 35:5282–5298, 2022. 2

[27] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 2023. 3

[28] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23023–23033, 2023. 7

[29] Liang Peng, Shuangji Yang, Yi Bin, and Guoqing Wang. Progressive graph attention network for video question answering. In *ACM Multimedia*, pages 2871–2879. ACM, 2021. 2

[30] Min Peng, Chongyang Wang, Yuan Gao, Yu Shi, and Xiang-Dong Zhou. Multilevel hierarchical network with multiscale sampling for video question answering. In *IJCAI*, pages 1276–1282. ijcai.org, 2022. 2

[31] 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 *ICML*, pages 8748–8763, 2021. 7

[32] Rajkumar Rajavel, Sathish Kumar Ravichandran, Karthikeyan Harimoorthy, Partheeban Nagappan, and Kanagachidambaresan Ramasubramanian Gobichettipalayam. Iot-based smart healthcare video surveillance system using edge computing. *Journal of ambient intelligence and humanized computing*, pages 1–13, 2022. 1

[33] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A local-to-global approach to multi-modal movie scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10146–10155, 2020. 1

[34] Ahjeong Seo, Gi-Cheon Kang, Joonhan Park, and Byoung-Tak Zhang. Attend what you need: Motion-appearance synergistic networks for video question answering. In *ACL/IJCNLP (1)*, pages 6167–6177. Association for Computational Linguistics, 2021. 2

[35] Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, Shayne Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, Tu Vu, Yuexin Wu, Wuyang Chen, Albert Webson, Yunxuan Li, Vincent Zhao, Hongkun Yu, Kurt Keutzer, Trevor Darrell, and Denny Zhou. Flan-moe: Scaling instruction-finetuned language models with sparse mixture of experts. *CoRR*, abs/2305.14705, 2023. 2

[36] Mohammad Shorfuzzaman, M Shamim Hossain, and Mohammed F Alhamid. Towards the sustainable development of smart cities through mass video surveillance: A response to the covid-19 pandemic. *Sustainable cities and society*, 64: 102582, 2021. 1

[37] Amit Singh, Albert Haque, Alexandre Alahi, Serena Yeung, Michelle Guo, Jill R Glassman, William Beninati, Terry Platchek, Li Fei-Fei, and Arnold Milstein. Automatic detection of hand hygiene using computer vision technology. *Journal of the American Medical Informatics Association*, 27(8):1316–1320, 2020. 1

[38] Mattia Soldan, Mengmeng Xu, Sisi Qu, Jesper Tegner, and Bernard Ghanem. Vlg-net: Video-language graph matching network for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3224–3234, 2021. 1

[39] Yale Song, Miriam Redi, Jordi Vallmitjana, and Alejandro Jaimes. To click or not to click: Automatic selection of beautiful thumbnails from videos. In *CIKM*, 2016. 7

[40] G Sreenu and Saleem Durai. Intelligent video surveillance: a review through deep learning techniques for crowd analysis. *Journal of Big Data*, 6(1):1–27, 2019. 1

[41] Reuben Tan, Bryan Plummer, Kate Saenko, Hailin Jin, and Bryan Russell. Look at what i'm doing: Self-supervised spatial grounding of narrations in instructional videos. *Advances in Neural Information Processing Systems*, 34:14476–14487, 2021. 1

[42] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, pages 4631–4640. IEEE Computer Society, 2016. 2

[43] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016. 1

[44] 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. 2

[45] Waseem Ullah, Amin Ullah, Ijaz Ul Haq, Khan Muhammad, Muhammad Sajjad, and Sung Wook Baik. Cnn features with bi-directional lstm for real-time anomaly detection in surveillance networks. *Multimedia tools and applications*, 80:16979–16995, 2021. 1

[46] Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. All in one: Exploring unified video-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6608, 2023. 7

[47] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun

Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 6, 7, 1

[48] Yuxuan Wang, Zilong Zheng, Xueliang Zhao, Jinpeng Li, Yueqian Wang, and Dongyan Zhao. VSTAR: A video-grounded dialogue dataset for situated semantic understanding with scene and topic transitions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5036–5048, Toronto, Canada, 2023. Association for Computational Linguistics. 1

[49] Bo Wu, Shoubin Yu, Zhenfang Chen, Josh Tenenbaum, and Chuang Gan. STAR: A benchmark for situated reasoning in real-world videos. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. 1, 3

[50] Junbin Xiao, Xindi Shang, Xun Yang, Sheng Tang, and Tat-Seng Chua. Visual relation grounding in videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 447–464. Springer, 2020. 1

[51] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, pages 9777–9786, 2021. 1, 3, 6

[52] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 2, 3

[53] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Fei Huang, and Jingren Zhou. mplug-2: A modularized multi-modal foundation model across text, image and video. In *ICML*, pages 38728–38748. PMLR, 2023. 6, 7, 1

[54] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, pages 1666–1677. IEEE, 2021. 2, 7

[55] Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. Hitea: Hierarchical temporal-aware video-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15405–15416, 2023. 7

[56] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 2023. 2, 6, 7, 1

[57] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019. 1, 3, 6

[58] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *CVPR*, pages 8807–8817, 2019. 3

[59] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5831–5840, 2018. 2, 4

[60] Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: Datasets, algorithms and challenges. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6439–6455, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. 1

Towards Fine-Grained Video Question Answering

Supplementary Material

A. Details of Scene Graph Predictor.

In this section, we detail the process of the scene graph predictor. Specifically, with object bounding boxes B, the object context $\mathbf{C}$ is first generated using a bidirectional LSTM layer:

$$\mathbf{C} = \text{biLSTM}([\mathbf{f}_i; \mathbf{W}_{ctx}\mathbf{p}_i]_{i=1,\dots,n}) \tag{10}$$

where $\mathbf{W}_{ctx}$ is a learnable matrix. We then use a biLSTM layer and an MLP layer to encode each object into edge contexts:

$$\hat{\mathbf{o}}_i = \text{argmax}(\mathbf{W}_o\text{LSTM}([\mathbf{c}_i; \hat{\mathbf{o}}_{i-1}]))$$
$$\mathbf{D} = \text{MLP}(\text{biLSTM}([\mathbf{c}_i; \mathbf{W}_d\hat{\mathbf{o}}_{i-1}]))$$

where $\mathbf{W}_d, \mathbf{W}_o$ are learnable matrices. In the end, for each pair of objects $(\mathbf{d}_i, \mathbf{d}_j)$, the scene graph feature $\mathbf{s}_{i,j}$ and the probability $\Pr(x_{i\to j})$ is generated:

$$\mathbf{s}_{i,j} = (\mathbf{W}_h\mathbf{d}_i)(\mathbf{W}_t\mathbf{d}_j)$$
$$\Pr(x_{i\to j}) = \text{softmax}(\mathbf{W}_r\mathbf{s}_{i,j})$$

Finally, top k filtering is performed so that only the features with the top k probabilities $\mathbf{S} = \{\mathbf{s}_i, \dots, \mathbf{s}_k\}$ are saved for the next stage. The scene graph predictor is trained beforehand and kept frozen during the VideoQA training.

B. Calculation of WUPS@0.9

With evaluation dataset $\mathbf{Q}$, the WUPS@0.9 score of the prediction $\hat{q}$ with respect to ground truth q is given by

$$WUPS = \frac{1}{|Q|}\sum_{q\in Q}\min\{\prod_{q_i\in A}\max_{\hat{q}_j} W_\gamma(q_i, \hat{q}_j), \prod_{\hat{q}_i}\max_{q_j} W_\gamma(\hat{q}_i, q_j)\}$$

and W_γ is given by

$$W_\gamma(q_i, \hat{q}_j) = \begin{cases} W(q_i, \hat{q}_j) & \text{if } W(q_i, \hat{q}_j) \geq \gamma \\ 0.1W(q_i, \hat{q}_j) & \text{if } W(q_i, \hat{q}_j) < \gamma \end{cases}$$

where we take $\gamma = 0.9$ to calculate WUPS@0.9.

C. Experimental Setup

We evaluate the models on MOMA-QA in both fine-tuned and zero shot settings. The details of each experiment are included below.

C.1. Zero Shot

We evaluate the performance of current state-of-the-art models on MOMA-QA in a zero-shot context. The experiment includes closed vocabulary models such as InternVideo [47] and mPLUG-2 [53], as well as open vocabulary models like BLIP-2 [21] and SeViLa [56]. Each model is assessed on the test split of MOMA-QA employing their respective optimal pre-trained parameters. For closed vocabulary models, answers are matched to the nearest word in the model's vocabulary when the precise answer falls outside its predefined vocabulary.

For our model, SGVLM, the scene graph predictor is trained on the scene graph dataset Visual Genome [50]; the frame localizer is trained on QVHighlights; and the full model is trained on NExT-QA [51] before being evaluated on MOMA-QA.

C.2. Fine-tuned

We evaluate the same baseline models with our model on MOMA-QA in a fine-tuned setting. We use the same initial weight as we used in the Zero Shot Experiment. Each model is trained on one computation node with four NVIDIA A6000s for a maximum of 5 epochs using the default hyperparameter settings from each model. The performance on the test dataset is reported.

For our model, we use the same starting point as the zero shot experiment. The scene graph predictor and vision backbone are tuned on MOMA-QA. The frame localizer is also tuned while training the full model for VideoQA.

C.3. Experiments on Public Datasets

In addition, we also evaluate models on NeXT-QA, a public dataset for video question answering, and QVHighlights, a public dataset for joint moment retrieval and highlight detections. NeXT-QA provides both multiple-choice and open-ended questions. For ease of comparison with baselines, we use the multiple-choice version of the dataset. Specifically, during evaluation, we take the probabilities for letters A, B, C, and D respectively, and choose the one with the highest probability as the prediction. This follows the standard practice employed in SeViLA [56] and eliminates the probability that the model predicts unrelated tokens.

For QVHighlights, only the frame retriever component of SGVLM is trained and evaluated. After tuning the model on the validation dataset, we train the model on a joint train-validation dataset and evaluate the model on the hidden test split.

D. Effect of Attention Masks

Model	Moment Retrieval					HD	
	R1		mAP			$\geq$ Very Good	
	@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
Without Mask	63.81	47.35	62.21	42.32	40.60	39.69	63.55
With Mask	**64.65**	**48.06**	**63.12**	**43.19**	**41.13**	**40.17**	**64.19**

Table 7. **Ablation Results on QVHighlights Validation Split.** The best in each column is bolded.

Table 7 shows the effect of attention masks on the performance of the frame localizer. As shown in the table, the variant with the attention mask achieves a higher score on all metrics, with up to 0.91% advantage on mAP@0.5. These results demonstrate the effectiveness of attention masks on the multi-modality input of the frame localizer.