

REVEAL: Relation-based Video Representation Learning for Video-Question-Answering

Sofian Chaybouti^{1,2}

¹Goethe University Frankfurt

Walid Bousselham^{1,2}

²Tuebingen AI Center/University of Tuebingen

Moritz Wolter³

³University of Bonn

Hilde Kuehne^{1,2,4}

⁴MIT-IBM Watson AI Lab

Abstract

Video-Question-Answering (VideoQA) comprises the capturing of complex visual relation changes over time, remaining a challenge even for advanced Video Language Models (VLM), i.a., because of the need to represent the visual content to a reasonably sized input for those models. To address this problem, we propose RElation-based Video rEpresentAtion Learning (REVEAL) a framework designed to capture visual relation information by encoding them into structured, decomposed representations. Specifically, inspired by spatiotemporal scene graphs, we propose to encode video sequences as sets of relation triplets in the form of (subject-predicate-object) over time via their language embeddings. To this end, we extract explicit relations from video captions and introduce a Many-to-Many Noise Contrastive Estimation (MM-NCE) together with a Q-Former architecture to align an unordered set of video-derived queries with corresponding text-based relation descriptions. At inference, the resulting Q-former produces an efficient token representation that can serve as input to a VLM for VideoQA.

We evaluate the proposed framework on five challenging benchmarks: NeXT-QA, Intent-QA, STAR, VLEP, and TVQA. It shows that the resulting query-based video representation is able to outperform global alignment-based CLS or patch token representations and achieves competitive results against state-of-the-art models, particularly on tasks requiring temporal reasoning and relation comprehension. The code and models will be publicly released upon acceptance.

1. Introduction

Videos capture rich sets of information, including the static visual information of a scene and the dynamic evolution of actors, objects, and their relationships over time. Understanding these complex spatiotemporal relations poses a significant challenge for current video understanding systems, as all those aspects need to be represented efficiently. One of the main tasks in this context is the problem of VideoQA [35, 67, 69, 76]. Approaches that do well here usually rely on pre-trained vision-language image backbones like CLIP

[49] and BLIP2 [32], processing videos by extracting frame representations and combining these with Large Language Models (LLMs) [26, 42]. However, these models struggle with object relations [37, 77], action detection [4, 37, 44, 66], and compositional understanding [4, 37], issues that are exacerbated with temporal sequences. While recent works have shown that LLMs can compensate those limitations via strong language priors [26, 35, 42, 60], image- and video-language approaches still mostly rely on global video-text alignment representations to encode the video input.

To address this problem, we propose RElation-based Video rEpresentAtion Learning (REVEAL). This framework learns video representations by explicitly modeling the content as object relations over time via relation triplets in the form of (subject-predicate-object). Our relation-based approach is inspired by prior work from video scene graphs context [9, 22, 51, 58]. However, scene graphs usually encode triplets via class indices, limiting the setting to close-ended and hand-annotated small-scale scenarios and hindering scalability. Inspired by this, REVEAL seeks to leverage this representation to learn general open-ended and web-supervised representations for video data.

To achieve this, we first leverage LLMs to convert captions into one or more relation triplets, allowing us to source triplets at scale. The resulting triplets can be considered minimum viable sentences, allowing a standard text encoding, e.g., by a sentence encoder, resulting in one embedding representation per triplet and J relation embeddings to describe a particular video. On the video side, we leverage a Q-Former architecture to encode the visual representation of one or more frames into a fixed set of vision queries. To train the Q-Former, we must match the fixed number of unordered vision queries to a variable number of unordered text triplet representations. To address this problem, we propose a Many-to-Many Noise Contrastive Estimation (MM-NCE) loss formulation, which aligns two sets of matching but unordered, incomplete sets, e.g., in our case, vision-based queries with corresponding text-based relation embeddings. Practically, MM-NCE maximizes the similarity between matched query-relation pairs while contrasting them against all unmatched pairs. This allows us to train the Q-former

so that the resulting query tokens approximate the relation encodings of the video. The resulting vision queries can then be used to fine-tune a standard VLM architecture to address video-language-related tasks such as VideoQA.

We evaluate REVEAL on five VideoQA datasets, NeXT-QA, Intent-QA, STAR, VLEP, and TVQA, demonstrating competitive performance compared to state-of-the-art methods. It shows that query-based representations, empowered by MM-NCE, are particularly effective at connecting video and text when adapting to LLMs. Our analysis further reveals that initializing the relation encoder with a contrastively trained sentence embedder significantly enhances semantic alignment compared to alternatives like CLIP’s text encoder. We summarize the contributions of this work as follows:

- We propose a new encoding for web-based video learning by modeling relations in videos as target representation.
- We propose a MM-NCE loss for contrastive learning over two sets of matching but unordered, incomplete sets.
- We provide an extensive evaluation showing the efficacy of query-based representations and the role of MM-NCE in the context of state-of-the-art VideoQA architectures.

2. Related Work

VLMs for VideoQA Video understanding, particularly VideoQA, has witnessed significant advancement with the emergence of LLMs and Large Vision-Language Models. Early approaches to VideoQA emerged in response to increasingly challenging benchmarks designed to test various aspects of video understanding. The complexity of VideoQA as a task is evidenced by the diverse set of benchmarks, each targeting different reasoning capabilities: TVQA [28] challenged models with understanding TV show content requiring integration of visual cues and dialogue; STAR [67] focused on situated reasoning about object interactions in indoor environments; NextQA [69] emphasized causal and temporal reasoning across everyday activities; IntentQA [33] specifically tested models’ ability to understand human intentions and motivations behind observed actions; and VLEP [29] evaluated models’ capacity to predict future events based on observed video content.

In addressing these challenges, early approaches predominantly treated VideoQA as a classification task, where video and question features were fed into classification layers to select from a fixed set of answer choices [11, 21, 71]. These methods typically employed CNN-RNN architectures, attention mechanisms, or memory networks to capture temporal dynamics, but their classification-based paradigm fundamentally limited their reasoning capabilities and prevented them from leveraging the generative power and world knowledge inherent in modern LLMs. Graph-based approaches like SHG-VQA [58] and VGT [70] attempted to model explicit relations between objects but remained constrained by closed-vocabulary limitations, small-scale datasets, and the

classification-based framework. These methods struggled with reasoning tasks due to a lack of semantic understanding.

Recent approaches have explored the direct application of VLMs to videos. IG-VLM [25] represents videos as image grids, while SLOWFAST-LLaVA [72] employs multi-scale temporal pooling for feature extraction. While effective for general understanding, these methods often struggle with complex temporal reasoning, which REVEAL addresses through explicit relation modeling. Further, the success of instruction-tuning in image-LLM connections [5, 16, 39, 68] has inspired similar approaches for video understanding. Video-ChatGPT [42], VideoChat [34], and their successors VideoChat2 [35] and VideoGPT+ [41] focus on video-conversation capabilities. Notable advances include Video-Llama [79]’s multi-modal processing, Video-LLaVA [36]’s unified representation space, and MotionEpic [12]’s “Video-of-Thought” framework. Llama-VQA [26] and Vamos [62] finetune adapters specifically for VideoQA. LLaVA-Next-Interleave [31], MPLUG-OWL-3 [73], and LLaVA-One Vision [30] have further advanced instruction-tuning approaches with powerful vision backbones. Finally, recent works have focused on unsupervised frame selection for this task, like Sevilla, Vila, and LVNet [46, 63, 75]. These approaches use large vision backbones and Gumbel-Softmax [20] to discriminate frames, achieving strong results with a handful of frames. While orthogonal to REVEAL’s relation-based approach, future work could combine these methods for more efficient videoQA.

Video-Language Pretraining Video-language pretraining has evolved significantly, with diverse architectural paradigms emerging to address the challenges of temporal modeling and multimodal alignment. Several key approaches have shaped this landscape: Q-former-based architectures like BLIP-2 [24] and its video adaptations [14, 36] use query-based cross-attention to bridge vision and language models; encoder-decoder frameworks like InternVideo [64] and InternVideo2 [65] combine masked video modeling with video-language contrastive learning; and unified architectures such as All-in-One [2] employ “token rolling” for efficient temporal modeling. Contrastive learning approaches have been particularly influential, with works like FrozenBiLM [3], VideoCLIP [17], CLIP4Clip [19] and CLIP2Video [13], establishing effective video-text alignment techniques. Temporal modeling has been addressed through hierarchical approaches in HiTeA [48] and HERO [38], while UniVL [18] pioneered joint understanding and generation objectives. Recent advances include VidL [7], which presents a progressive recipe for video-language model construction, and MERLOT [78], which leverages YouTube transcripts for self-supervised learning.

3. RElation-based Video rEpresentAtion Learn-ing (REVEAL)

REVEAL is a framework designed to capture visual relation information in videos by encoding them into structured, decomposed representations. This section is structured as follows: Sec. 3.1 describes the relation triplet sourcing from video captions, Sec. 3.2 the overall architecture of REVEAL, Sec. 3.3 details the relation modeling, Sec. 3.4 describes the MM-NCE loss for aligning unordered sets of relations, and Sec. 3.5 covers the implementation details.

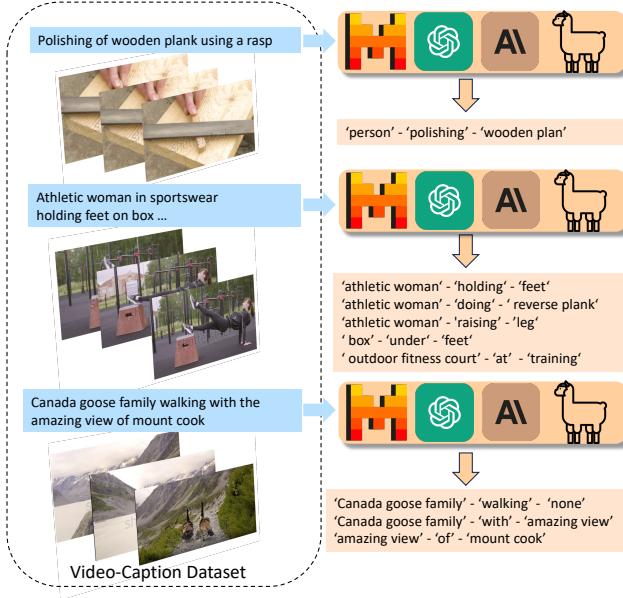


Figure 1. Relation extraction pipeline: Mistral-7B decomposes WebVid-2M captions into (subject-predicate-object) triplets.

3.1. Relation Extraction from Video Captions

We develop a relation extraction pipeline to transform natural language video captions into structured relation triplets (*subject-predicate-object*). Traditional approaches often depend on manually annotated datasets or rule-based methods, limiting scalability [53, 54, 56, 76]. Compared to that, REVEAL leverages the Mistral-7B model [23] to automate and scale extraction from large-scale datasets like WebVid-2M [3]. To guide the LLM in decomposing unstructured captions into meaningful relation triplets, we in-context learning, detailed in Supplement Sec. 10, to identify and extract relevant relations. This pipeline automatically generates multiple relation triplets per video, as illustrated in Figure 1, providing a decomposed representation of the video caption respective to the visual content.

3.2. REVEAL Architecture

Our approach represents videos as sets of relation triplets in the form of *subject-predicate-object*. Unlike methods relying on finite indexed triplets [58] or separate object-predicate

classification [15, 52], we learn relation representations from language embeddings by aligning video-derived queries with text-derived relation embeddings.

As shown in Figure 2, the REVEAL architecture consists of four main components: (1) a vision encoder to compute frame-level features via a pretrained backbone; (2) a temporal encoder to capture the temporal dependencies across features from different frames; (3) a Relation Q-Former to transform the resulting visual features into vision queries; and (4) a Relation Encoder to encode text-based relation triplets for supervision. During training, our MM-NCE loss aligns the vision queries with relation embeddings through Hungarian matching followed by contrastive learning, optimizing all components except the frozen vision backbone.

3.3. Relation Modeling

For a video \mathcal{V} , we begin with transforming the video into visual tokens. A visual encoder $f(\cdot)$ processes each video's frames, producing a set of features: $(\mathbf{x}_n)_{n \in \{1..N\}} = f(\mathcal{V})$, where N denotes the number of tokens per video. These tokens serve as input for relation modeling.

Relation Q-former: To transform learnable queries $(\mathbf{v}_m^0)_{m \in \{1..M\}}$ into vision queries $(\mathbf{v}_m)_{m \in \{1..M\}}$, we employ a Q-former architecture [6]. This module performs cross-attention between the initial queries and the video's visual tokens $(\mathbf{x}_n)_{n \in \{1..N\}}$:

$$(\mathbf{v}_m)_{m \in \{1..M\}} = g((\mathbf{v}_m^0)_{m \in \{1..M\}}, (\mathbf{x}_n)_{n \in \{1..N\}}),$$

The resulting vision queries are processed through a feed-forward network to yield relation embeddings aligned with text-derived triplets.

Relation Encoder: In parallel, text relations $(\mathbf{t}_j)_{j \in \{1..J\}}$ associated with the video are passed through a text encoder $h(\cdot)$ to get relation embeddings $(\mathbf{r}_j)_{j \in \{1..J\}} = h((\mathbf{t}_j)_{j \in \{1..J\}})$. Practically, we leverage a pre-trained sentence embedder, initialized with contrastively trained models like Sentence-BERT [50].

Finally, the vision queries $(\mathbf{v}_m)_{m \in \{1..M\}}$ are aligned with the text-derived relation embeddings $(\mathbf{r}_j)_{j \in \{1..J\}}$ via the proposed MM-NCE loss.

3.4. Relation Loss Function: Many-to-Many Noise Contrastive Estimation

We introduce Many-to-Many Noise Contrastive Estimation (MM-NCE) as a contrastive learning approach designed to align unordered sets of relations. The key challenge is that relation triplets extracted from video captions form an unordered set with no predefined temporal correspondence to visual elements in the video. This presents two difficulties: the number of extracted relation triplets may differ from the number of visual queries, requiring a flexible matching strategy, and unlike traditional video-text alignment where a

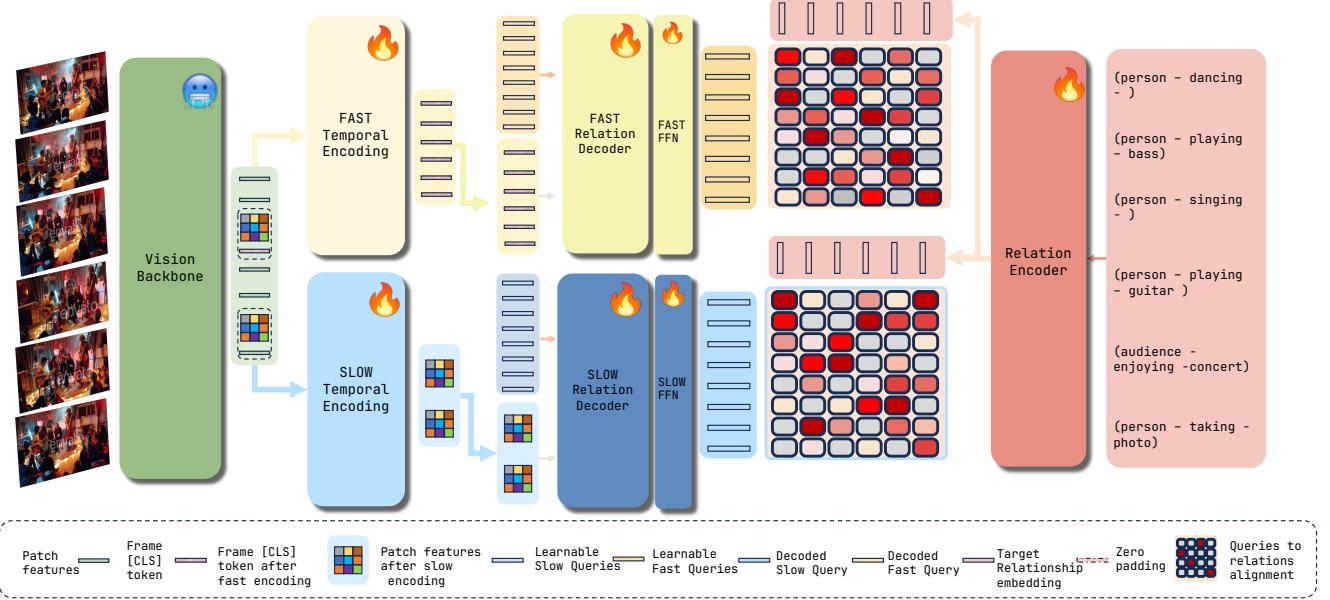


Figure 2. REVEAL architecture for relation-based video representation learning. The model processes videos through dual pathways: a Fast Pathway (16 frames) for global context and a Slow Pathway (4 frames) for spatial details. Key components include Vision Encoders (CLIP ViT), Temporal Encoders (transformers), Relation Q-formers, and a Relation Encoder (Sentence-RoBERTa). The training uses our MM-NCE loss to align vision queries with text-derived relation triplets.

single caption corresponds to an entire video, our approach must determine which specific vision query correspond to which relation embedding without explicit supervision. For a batch with samples $k \in \mathcal{B}$, we consider each video as \mathcal{V}_k , the text-derived relation embeddings as $(\mathbf{r}_j^{(k)})_{j \in \mathcal{J}^{(k)}}$, with $\mathcal{J}^{(k)} = \{1..J^{(k)}\}$, and the vision queries as $(\mathbf{v}_m^{(k)})_{m \in \{1..M\}}$. We first determine the optimal matching between them using Hungarian matching, therefore creating a set of query-relation positive pairs for each video in the batch:

$$\sigma^{(k)} = \operatorname{argmax}_{\sigma \in \mathcal{S}_{J^{(k)}, M}} \sum_{j \in \mathcal{J}^{(k)}} s_c(\mathbf{r}_j^{(k)}, \mathbf{v}_{\sigma(j)}^{(k)}), \quad (1)$$

where $\mathcal{S}_{J^{(k)}, M}$ represents the set of injective mappings from $\mathcal{J}^{(k)}$ to $\{1..M\}$ and with the cosine similarity

$$s_c(\mathbf{r}, \mathbf{v}) = \frac{\mathbf{r}^T \mathbf{v}}{\|\mathbf{r}\| \|\mathbf{v}\|}. \quad (2)$$

Note that, in equation 1, not all vision queries are paired to a text-derived relation embedding when $J^{(k)} < M$; the resulting mapping $\sigma^{(k)}(.)$ is injective but not surjective. This is a key property of our approach: it is designed to handle varying numbers of text relations per video. Eventually, only paired vision queries contribute to the loss defined below.

In the following equations, we omit the learnable parameters. The MM-NCE loss then consists of two symmetric

terms. L_q measures query-to-relation alignment:

$$L_{q \rightarrow r} = \sum_{k \in \mathcal{B}} \log \frac{\exp(s_c(\mathbf{r}_j^{(k)}, \mathbf{v}_{\sigma^{(k)}(j)}^{(k)}) / \tau)}{\sum_{j' \in \mathcal{J}^{(k)}} \exp(s_c(\mathbf{r}_j^{(k)}, \mathbf{v}_{\sigma^{(k)}(j')}^{(k)}) / \tau)}, \quad (3)$$

and for the relation-to-query alignment term $L_{r \rightarrow q}$:

$$L_{r \rightarrow q} = \sum_{k \in \mathcal{B}} \log \frac{\exp(s_c(\mathbf{r}_j^{(k)}, \mathbf{v}_{\sigma^{(k)}(j)}^{(k)}) / \tau)}{\sum_{k' \in \mathcal{B}} \exp(s_c(\mathbf{r}_j^{(k)}, \mathbf{v}_i^{(k')}) / \tau)}. \quad (4)$$

Here, k' and i index over all videos in batch \mathcal{B} and vision queries from a video, respectively, creating negative pairs from other videos. The temperature parameter τ is learnable.

$L_{q \rightarrow r}$ and $L_{r \rightarrow q}$ allow us to compute the MM-NCE-loss:

$$L_{\text{MM-NCE}} = L_{q \rightarrow r} + L_{r \rightarrow q}. \quad (5)$$

MM-NCE pulls matched query-relation pairs closer in embedding space while pushing negative pairs apart, handling the unordered nature of relations through Hungarian matching rather than requiring predefined correspondences. It specifically allows the handling of varying numbers of annotations per video. When some vision queries are not

matched to annotated relations, the model can freely learn to model relations in a video even when not annotated if they appear in other videos in the training data. Thus, it can also deal with non-exhaustive annotation. Unlike Multiple Instance Learning and Noise Contrastive Estimation (MIL-NCE) [43], designed to align multiple captions to a single representation, this approach enforces a one-to-one correspondence between the multiple video representations, *i.e.*, the vision queries, and the corresponding text relations.

3.5. Implementation Details

Slow-Fast Video Processing Following recent work [41, 72], REVEAL employs a dual-pathway architecture to capture both global context and fine-grained spatial information, enhancing relation understanding while balancing computational efficiency: the **Fast Pathway** uses [CLS] tokens across 16 frames for efficient temporal aggregation and high-level motion understanding; the **Slow Pathway** processes patch features from four carefully selected frames for detailed spatial information and object-level relationship modeling. Each pathway processes its respective features using a dedicated temporal encoder and relation Q-former. The temporal encoders model dependencies across frames, with the Fast pathway capturing global changes and long-range temporal dynamics and the Slow pathway specifically modeling patch relationships across frames for fine-grained spatial reasoning. The relation decoders perform cross-attention with visual features to transform learnable queries into relation embeddings representing meaningful subject-predicate-object triplets.

VideoQA Finetuning To evaluate REVEAL on multiple-choice VideoQA tasks, we adapt the frozen pre-trained video-derived relation features to LLMs, providing a decomposed representation for question answering, as illustrated in Figure 3. Following previous work [26, 62], our finetuning approach integrates REVEAL’s video relation embeddings into pre-trained LLMs using Llama adapters [80]. The process begins with REVEAL processing segmented videos (1–8 segments) in parallel, modeling 16 vision queries per segment (8 per pathway), which yields 16–128 embeddings per video. These embeddings are then projected into the LLM’s vocabulary space via a linear transformation. For temporal alignment, each group of 16 vision queries corresponds to its respective video segment, with special tokens distinguishing between Slow and Fast Pathway outputs and learnable temporal tokens encoding segment positions. Our training methodology follows Flipped-VQA [26], employing three complementary tasks: the main task ($VQ \rightarrow A$) predicts answers from video vision queries and questions, while auxiliary tasks predict questions from vision queries and answers ($VA \rightarrow Q$) and vision queries from questions and answers ($QA \rightarrow V$). This multi-task approach reduces reliance on linguistic bias and enhances visual grounding, with REVEAL

frozen to preserve the pre-trained representations.

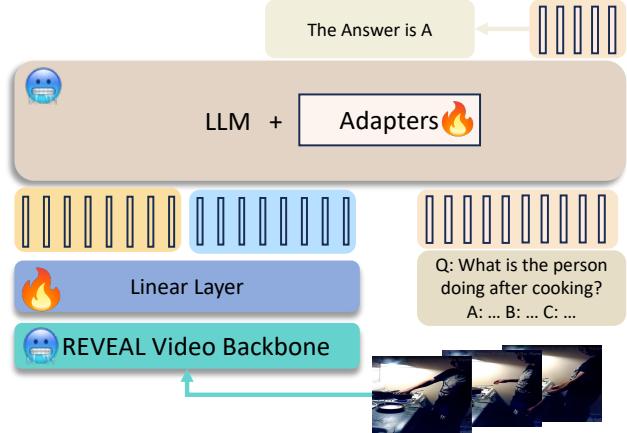


Figure 3. Overview of the VideoQA finetuning approach. The framework integrates pre-trained relation embeddings from our model with LLMs via adapters.

4. Experiments

4.1. Datasets

Pretraining Datasets: We pretrain REVEAL on the **WebVid-2M** dataset, a large-scale collection of **2.5 million** video-caption pairs sourced from public web platforms [3]. Relation triplets are extracted from captions using the **Mistral-7B** model [23]. Post-extraction, automated filtering removes ambiguous or redundant triplets, yielding an average of four relations per video. To enhance relation diversity and robustness, we incorporate annotations of 8k videos from **Charades** [54] and 3k videos from **VidOR** [53], splitting each video into clips with 4–8 relations, adding approximately 80k clips to the training set. To prevent data leakage, we ensure no selected clips from Charades or VidOR overlap with STAR, NeXT-QA, or Intent-QA evaluation sets. The relation extraction and filtering details are provided in the supplement section 10.

VideoQA Evaluation Datasets: We finetune and evaluate the resulting model on five diverse VideoQA benchmarks: ◇**STAR** [67] features 60K questions across four reasoning types (interaction, sequence, prediction, feasibility) with 22K indoor activity clips. Its procedurally generated questions require understanding object interactions and action consequences. ◇**NExT-QA** [69] contains 52K manually annotated QA pairs over 5,440 videos, categorized as causal (48%), temporal (29%), or descriptive (23%) testing event causation reasoning. ◇**Intent-QA** [33] extends NExT-QA with 16K QA pairs focused on intention understanding through four question types, challenging models to infer motives from observed actions. ◇**TVQA** [28] comprises 152K QA pairs from 21K TV show clips, with five-choice questions. Its dialogue-heavy content requires integrating visual and linguistic cues across narratives. ◇**VLEP** [29] presents a binary event pre-

Method	Specifications	Language Backbone	Vision Backbone	Int	Seq	Pred	Feas	All
SHG-VQA (val set) [58]	FT	BERT	SlowR50-K400	48.0	42.0	35.3	32.5	39.5
All-in-One [61]	PT + FT	All-in-One	All-in-One	47.5	50.8	47.7	44.0	47.5
InternVideo [64]	PT + FT	CLIP text encoder	ViT-H/14	62.7	65.6	54.9	51.9	58.7
Sevilla [75]	FT + FS	BLIP-2 (FlanT5-XL)	BLIP-2 (ViT-G/14)	63.7	70.4	63.1	62.4	64.9
ViLA [63]	FT + FS	BLIP-2 (FlanT5-XL)	BLIP-2 (ViT-G/14)	70.0	70.4	65.9	<u>62.2</u>	67.1
IG-VLM [25]	ZS	Llava 1.6	ViT-L/14	49.3	50.1	49.5	48.8	49.6
Llama-VQA [26] (baseline)	LLM-A	Llama1	ViT-L/14	<u>66.2</u>	<u>67.9</u>	<u>57.2</u>	<u>52.7</u>	<u>65.4</u>
REVEAL (ours)	PT + LLM-A	Llama1	ViT-L/14	60.0	<u>70.7</u>	72.5	<u>68.4</u>	67.9
Llama-VQA* [26] (baseline)	LLM-A	Llama3	ViT-L/14	59.8	<u>67.2</u>	59.8	50.4	65.4
REVEAL (ours)	PT + LLM-A	Llama3	ViT-L/14	<u>59.7</u>	70.8	70.7	68.7	67.5

Table 1. Performance comparison on STAR dataset for situated reasoning VideoQA across different question types (Interaction, Sequence, Prediction, and Feasibility). Specifications: PT = Pretraining, FT = Finetuning, FS = Frame Selection, ZS = Zero-Shot, LLM-A = LLM with Adapters. * indicates that we run the baseline evaluation ourselves.

Method	Specifications	Language Backbone	Vision Backbone	Caus	Temp	Des	All
All-in-One [61]	PT + FT	All-in-One	All-in-One	48.6	48.0	63.2	50.6
Video-Llama [79]	IT + ZS	Llama	ViT-G/14	57.4	59.2	72.3	60.6
VideoChat [35]	IT + ZS	StableVicuna	BLIP-2 (ViT-G/14)	61.5	63.5	82.1	61.8
HiTeA [74]	PT + FT	BERT-Base	MViT-Base	58.3	62.4	75.6	63.1
InternVideo [64]	PT + FT	CLIP text encoder	ViT-H	58.5	62.5	75.8	63.2
VideoChat2 [35]	IT + ZS	Llama1	UMT-L	64.7	68.7	76.1	68.6
LVNet [12]	ZS + FS	GPT-4o	GPT-4o	65.5	75.0	81.5	72.9
Sevilla [75]	FT + FS	BLIP-2 (FlanT5-XL)	BLIP-2 (ViT-G/14)	<u>69.4</u>	<u>74.4</u>	81.3	<u>73.8</u>
ViLA [63]	FT + FS	BLIP-2 (FlanT5-XL)	BLIP-2 (ViT-G/14)	71.4	73.6	<u>81.4</u>	74.8
IG-VLM [25]	VLM + ZS	Llava 1.6	ViT-L/14	63.1	57.3	74.9	63.1
SLOWFAST-LLava [72]	VLM + ZS	LLava-Next	ViT-L/14	–	–	–	64.2
Video-ChatGPT [42]	IT + ZS	LLaVA	ViT-L/14	64.1	66.9	75.7	64.4
Flipped-VQA (baseline) [25]	LLM + A	Llama1	ViT-L/14	72.7	69.2	75.8	72.0
REVEAL (ours)	PT + LLM-A	Llama1	ViT-L/14	<u>73.7</u>	<u>69.2</u>	76.5	72.7
REVEAL (ours)	PT + LLM-A	Llama3	ViT-L/14	75.3	69.9	78.5	74.0
Vamos [62]*	C + LLM-A	Llama3	ViT-L/14	76.1	73.7	80.4	76.0
REVEAL (ours)	C + PT + LLM-A	Llama3	ViT-L/14	77.8	74.4	81.9	<u>77.2</u>
Vamos [62]	C + LLM-A	Llama3	ViT-L/14	<u>77.2</u>	<u>75.3</u>	<u>81.7</u>	77.3
LLaVA-Next-Interleave [31]	IT + ZS	QWEN-1.5	SigLIP	–	–	–	77.9
MPLUG-OWL-3 [73]	IT	QWEN-2	SigLIP	–	–	–	<u>78.6</u>
LLaVA-One Vision [30]	IT	QWEN-2	SigLIP	–	–	–	79.4

Table 2. Performance comparison on NExT-QA dataset for causal, temporal, and descriptive VideoQA. *indicates reproduced results. diction task with 28K examples across 10K clips. Models must predict which of two events will occur next, testing anticipatory reasoning. Performance is measured by answer accuracy, with category breakdowns for STAR, NExT-QA, and Intent-QA highlighting task-specific strengths.

4.2. Training Details

Using CLIP’s ViT-L/14 [49] as the vision backbone, we process frame features with a two-layer transformer encoder followed by a 12-layer Q-former module for both the slow and the fast pathway output, resulting in eight learnable queries per pathway. This yields 16 vision query tokens per video clip, which are projected into the relation embedding space via a fully connected feed-forward network. The relation encoder is initialized with a pretrained sentence embedder ("all-

roberta-large-v1" from [55]) based on Sentence-BERT [50] and the RoBERTa-large architecture [40]. It transforms relation triplets, formatted as "Subject: *subj*, Predicate: *pred*, Object: *obj*", into single 1024-dimensional embeddings. Depending on the caption, one video can have multiple associated triplets. If more than eight text-derived relation embeddings are available, we randomly sample eight triplets. The resulting embedding sequence is further adapted with a one-layer feed-forward network.

We pretrain the model for five epochs on eight MI210 GPUs for approximately one day with the AdamW optimizer and a cosine-decayed learning rate of 5×10^{-5} . The resulting model comprises 590 million parameters.

We finetune the model for each benchmark separately. To this end, we follow best practices of previous works [26,

Method	Specifications	Language Backbone	Vision Backbone	CW	CH	TP&TN	All
HQGA [27]	FT	BERT	ResNeXt-101/ResNet-101	48.2	54.3	41.7	47.7
VGT [70]	FT	BERT	VGT	51.4	55.9	47.6	51.3
CaVIR [33]	FT	BERT	VGT	58.4	65.4	50.5	57.6
VideoChat [36]	IT + ZS	Stable Vicuna	BLIP-2 (ViT-G/14)	—	—	—	59.3
LVNet [12]	ZS + FS	GPT-4o	GPT-4o	75.0	74.4	62.1	71.7
Video-LLaVA [36]	IT + ZS	Vicuna-7B	ViT-L/14	—	—	—	62.5
Flipped-VQA* [26]	LLM-A	LLama3	ViT-L/14	73.7	72.6	57.3	69.5
REVEAL (ours)	PT + LLM-A	Llama3	ViT-L/14	74.0	77.4	66.8	72.8
Vamos [62]	C + LLM-A	Llama3	ViT-L/14	75.1	77.4	69.5	74.1
REVEAL (ours)	PT + C + LLM-A	Llama3	ViT-L/14	77.9	77.3	67.5	75.0

Table 3. Performance comparison on Intent-QA dataset for intention understanding through causal and temporal reasoning (CW: Causal Why, CH: Causal How, TP&TN: Temporal Previous & Next). * indicates that we run the baseline evaluation ourselves.

[62], keeping our pretrained video model, REVEAL, frozen and finetuning only a Linear layer and the Llama backbone via Llama-adapters [80] considering both Llama1 (7B) and Llama3 (8B)[10, 57] as our language models.

4.3. Comparison to State-of-the-Art Methods

We evaluate REVEAL against state-of-the-art methods across five VideoQA benchmarks.

STAR: Table 1 shows the comparison with state-of-the-art approaches on STAR. We improve by 2.5% compared to the Flipped-VQA baseline [26], with the same vision, language backbones, and finetuning setting. Furthermore, we achieve state-of-the-art results improving upon ViT-G/14-based ViLA [63] by 0.8% while using the significantly less powerful ViT-L/14. The most substantial gains appear in prediction (+6.6%) and feasibility (+6.2%) questions testing the understanding of interactions and temporal reasoning.

NExT-QA: On NExT-QA (Table 2), REVEAL with Llama3 achieves 74.0% accuracy and 72.7% with Llama1, outperforming the Flipped-VQA baseline (72.0%) using identical vision backbones. Additionally, we implement a REVEAL+Captioning baseline for comparison with Vamos[62], integrating off-the-shelf captioning to complement relation embeddings with text descriptions. This improves performance to 77.2% and is on par with Vamos’s results (77.3%) while outperforming our reproduced baseline by 1.2%.

Intent-QA: Table 3 shows REVEAL achieving 72.8% accuracy on Intent-QA, beating the Flipped-VQA baseline 3.3%. With complementary captions, REVEAL reaches 75.0%, surpassing Vamos (74.1%), with identical backbones.

TVQA and VLEP Datasets: On TVQA (Table 4), REVEAL achieves state-of-the-art performance (83.0%), outperforming Flipped-VQA (82.2%) by 0.8%. Similarly, on VLEP (Table 5), we achieve 73.5% surpassing Flipped-VQA by 2.5% and 1.2% with Llama1 and Llama3, respectively. These consistent improvements on datasets with different characteristics—from dialogue-heavy TV content to event prediction tasks—demonstrate the versatility and robustness of our relation-based approach.

Method	Specs.	Language	Vision	All
InternVid [64]	PT+FT	CLIP	ViT-H	57.2
Merlot [78]	PT+FT	RoBERTa	ResNet-50	78.7
VidL [7]	PT+FT	BERT	ViT-B/16	79.0
FrozenBiLM [3]	PT+ZS	DeBERTa	ViT-L/14	82.0
Flipped-VQA [25]	LLM-A	Llama1	ViT-L/14	82.2
REVEAL	PT+LLM-A	Llama3	ViT-L/14	83.0

Table 4. Performance on TVQA dataset.

Method	Specs.	Language	Vision	All
InternVideo [64]	PT+FT	CLIP	ViT-H	63.9
Merlot [78]	PT+FT	RoBERTa	ResNet-50	68.4
VideoChat [35]	IT+ZS	Stable Vicuna	ViT-G/14	62.0
SeViLA [75]	FT+FS	FlanT5-XL	ViT-G/14	68.9
ViLA [63]	FT+FS	FlanT5-XL	ViT-G/14	69.6
Video-LLaVA [36]	IT+ZS	Vicuna-7B	ViT-L/14	65.8
Flipped-VQA [25]	LLM-A	Llama1	ViT-L/14	71.0
Flipped-VQA* [25]	LLM-A	Llama3	ViT-L/14	72.3
REVEAL	PT+LLM-A	Llama3	ViT-L/14	73.5

Table 5. Performance on VLEP dataset. * indicates that we run the baseline evaluation ourselves.

4.4. Ablation Studies

We conduct respective ablation studies to validate the key components of REVEAL. Results are summarized in Tables 6 and 7 across STAR, NeXT-QA, and Intent-QA.

Video-Relation vs. Video-Caption Alignment: Table 6.a tests relation modeling with MM-NCE loss compared to caption-based NCE supervision. It shows that pretraining with relations and MM-NCE loss yields significant improvements (STAR: +33.9%, NeXT-QA: +16.5%, Intent-QA: +9.6%) over captions with standard NCE. These substantial gains validate the hypothesis that decomposing videos into structured relation triplets creates more effective representations than deriving representations from global captions.

Trainable vs. Frozen Relation Encoder: Table 6.b provides first a baseline for the proposed matching loss, optimizing the query representation by computing the best matches and later optimizing them via MSE. Second, we provide results for the same setup but with the proposed MM-NCE loss func-

Ablation	STAR					NeXT-QA				Intent-QA			
	In	Seq	Pre	Feas	All	C	T	D	All	CW	CH	TN	All
a) Annotations:													
Captions + NCE loss	32.1	35.2	28.7	29.6	31.5	58.4	56.1	49.7	56.3	69.2	63.5	50.8	61.2
relations + MM-NCE loss	58.4	65.6	69.1	68.4	65.4	74.0	68.3	77.7	72.8	74.9	74.0	62.5	70.8
b) Rel. Enc.:													
Frozen + MSE loss	59.3	68.9	75.2	70.6	66.4	73.1	66.6	73.5	71.1	73.9	73.9	55.5	68.9
Frozen + MM-NCE loss	59.7	69.0	73.1	69.8	67.9	73.8	68.8	76.2	72.6	73.7	74.4	63.1	71.4
Trainable + MM-NCE loss	61.4	69.3	75.0	72.0	69.4	75.3	69.9	78.5	74.0	74.6	75.5	65.6	71.8
c) LLM's video input:													
Without FFN layer	54.6	61.0	64.7	67.8	62.0	73.3	68.1	76.3	72.1	72.8	74.3	56.9	70.0
With FFN layer	61.4	69.3	75.0	72.0	69.4	75.3	69.9	78.5	74.0	74.6	75.5	65.6	71.8
d) Pathways:													
Slow	62.1	68.9	74.2	70.2	68.9	73.0	68.2	76.1	71.9	74.0	73.3	60.6	70.3
Fast	57.5	65.5	68.1	69.6	65.2	73.7	68.4	77.5	72.6	73.1	74.2	66.3	71.1
Slow-Fast	61.4	69.3	75.0	72.0	69.4	75.3	69.9	78.5	74.0	74.6	75.5	65.6	71.8

Table 6. a) Pretraining on relations compared to training on captions. Both models were pre-trained on WebVid only. The caption model was contrastively trained by attention pooling on the vision queries. b) Ablation on the trainable relation encoder c) Results of using the vision queries compared to the last hidden states from REVEAL. d) Ablation on the slow-fast architecture.

	STAR	NeXT-QA	Intent-QA
a) Initialization:			
Random init	68.3	71.0	69.2
RoBERTa-large	68.0	72.2	71.0
CLIP text encoder	68.5	72.3	71.4
Sentence embedder	69.4	74.0	71.8
b) relations:			
1	65.3	72.4	70.5
2	68.3	72.9	70.7
4	68.0	73.1	71.3
8	69.4	74.0	71.8

Table 7. Ablation on a) the initialization of the relation encoder and b) the number of relations used as input to the LLM.

tion. In both cases, the sentence embedder is kept frozen to evaluate the direct impact of the loss function, showing that MM-NCE provides better performance than matching followed by MSE. Third, we copy the second setup and make the sentence embedding trainable. A trainable encoder with MM-NCE loss consistently outperforms both a frozen encoder with MM-NCE (+1.5% on STAR) and a frozen encoder with MSE loss (+3.0% on STAR). This demonstrates that MM-NCE not only aligns relation sets but also enables the relation encoder to adapt to video-specific patterns, enhancing the semantic richness of our relation representations.

Vision Queries vs. Hidden States: We further evaluate the optimal input for the LLM. Namely, Table 6.c compares the results for using tokens before and after the FFN layer. This is motivated by the fact that the last layer in self-supervised learning can overfit on the objective. It shows that in our case, the output of the FFN projection outperforms the intermediate output of the Q-Former (STAR: +7.4%, NeXT-QA:

+1.2%, Intent-QA: +1.8%), confirming that explicitly modeling structured relations provides LLMs with interpretable and actionable representations.

Slow-Fast Architecture: Table 6.d assess the impact of the dual-pathway architecture. It shows that using both representations consistently outperforms single-pathway variants (STAR: +0.5% over Slow, +4.2% over Fast), showing that modeling relations can be improved by detailed spatial information for object identification and efficient temporal modeling for action recognition.

Relation Encoder Initialization: Table 7.a demonstrates that initializing the relation encoder with a contrastively trained sentence embedder significantly outperforms alternatives (e.g., +0.9% over CLIP on STAR). This supports our claim that effective relation modeling requires semantically rich embeddings that can discriminate between similar but distinct relations (e.g., "person opens door" vs. "person closes door"), which contrastive training naturally provides.

Number of Vision Queries: Table 7.b shows that increasing from 1 to 8 relations per pathway consistently improves performance (STAR: +4.1%, NeXT-QA: +1.6% and Intent-QA: +1.3%), validating our modeling approach. This confirms that videos are better represented as sets of multiple relations rather than single global entities, with each additional relation contributing meaningful information.

5. Conclusion

We presented REVEAL, a framework advancing video understanding through relation-based representation learning. By modeling videos as relation triplet sets and introducing MM-NCE loss for aligning unsorted relations, our approach creates structured embeddings that connect effectively with LLMs. Experiments show that decomposed relation-based representations outperform global alignment ones.

References

- [1] Alexander Aizman and Myle Ott. Webdataset. <https://github.com/webdataset/webdataset>, 2021. 1
- [2] Wang Alex, Jinpeng Ge Yixiao, Yan Rui, Ge Yuying, Lin Xudong, Cai Guanyu, Wu Jianping, Shan Ying, Qie Xiaohu, and Shou Mike, Zheng. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*, 2022. 2
- [3] Max Bain, Arsha Nagrani, GÜl Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 2, 3, 5, 7
- [4] Hritik Bansal, Yonatan Bitton, Idan Szpektor, Kai-Wei Chang, and Aditya Grover. Videocon: Robust video-language alignment via contrast captions. In *CVPR*, 2024. 1
- [5] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 2
- [6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*. Springer, 2020. 3
- [7] Feng Cheng, Xizi Wang, Jie Lei, David Crandall, Mohit Bansal, and Gedas Bertasius. Vindlu: A recipe for effective video-and-language pretraining. In *CVPR*, 2023. 2, 7
- [8] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2023. 1
- [9] Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *ICCV*, 2021. 1
- [10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 7
- [11] 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*, 2019. 2
- [12] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *ICML*, 2024. 2, 6, 7
- [13] Fang Han, Xiong Pengfei, Xu Luhui, and Chen Yu. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 2
- [14] Zhang Hang, Li Xin, and Bing Lidong. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 2
- [15] Roei Herzig, Alon Mendelson, Leonid Karlinsky, Assaf Arbel, Rogerio Feris, Trevor Darrell, and Amir Globerson. Incorporating structured representations into pretrained vision & language models using scene graphs. *EMNLP*, 2023. 3
- [16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ICLR*, 2022. 2
- [17] Xu Hu, Ghosh Gargi, Huang Po-Yao, Okhonko Dmytro, Aghajanyan Armen, Metze Florian, Zettlemoyer Luke, and Feichtenhofer Christoph. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 2
- [18] Luo Huaishao, Ji Lei, Shi Botian, Huang Haoyang, Duan Nan, Li Tianrui, Li Jason, Bharti Taroon, and Zhou Ming. Univil: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 2
- [19] Luo Huaishao, Ji Lei, Zhong Ming, Chen Yang, Lei Wen, Duan Nan, and Li Tianrui. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 2
- [20] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *ICLR*, 2017. 2
- [21] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2017. 2
- [22] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *CVPR*, 2020. 1
- [23] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 3, 5
- [24] Li Junnan, Li Dongxu, Xiong Caiming, and Hoi Steven. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *ICML*, 2022. 2
- [25] Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. An image grid can be worth a video: Zero-shot video question answering using a vlm. *arXiv preprint arXiv:2403.18406*, 2024. 2, 6, 7
- [26] Dohwan Ko, Ji Soo Lee, Wooyoung Kang, Byungseok Roh, and Hyunwoo J Kim. Large language models are temporal and causal reasoners for video question answering. *EMNLP*, 2023. 1, 2, 5, 6, 7
- [27] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for multimodal video question answering. *IJCV*, 2021. 7
- [28] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *EMNLP*, 2018. 2, 5
- [29] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. What is more likely to happen next? video-and-language future event prediction. *EMNLP*, 2020. 2, 5
- [30] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *TMLR*, 2025. 2, 6

- [31] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zeyun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *ICLR*, 2025. 2, 6
- [32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 1
- [33] Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. Intentqa: Context-aware video intent reasoning. In *ICCV*, 2023. 2, 5, 7
- [34] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 2
- [35] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. *CVPR*, 2024. 1, 2, 6, 7
- [36] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *EMNLP*, 2024. 2, 7
- [37] Kun-Yu Lin, Henghui Ding, Jiaming Zhou, Yu-Ming Tang, Yi-Xing Peng, Zhilin Zhao, Chen Change Loy, and Wei-Shi Zheng. Rethinking clip-based video learners in cross-domain open-vocabulary action recognition. *arXiv preprint arXiv:2403.01560*, 2024. 1
- [38] Li Linjie, Chen Yen-Chun, Cheng Yu, Gan Zhe, Yu Licheng, and Liu Jingjing. Hero: Hierarchical encoder for video+language omni-representation pre-training. *EMNLP*, 2020. 2
- [39] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 2
- [40] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 6
- [41] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*, 2024. 2, 5
- [42] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *ACL*, 2024. 1, 2, 6
- [43] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In *CVPR*, 2020. 5
- [44] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. Verbs in action: Improving verb understanding in video-language models. In *ICCV*, 2023. 1
- [45] Neptune team. neptune.ai: Metadata store for mlops, 2019. 1
- [46] Jongwoo Park, Kanchana Ranasinghe, Kumara Kahatapitiya, Wonjeong Ryoo, Donghyun Kim, and Michael S Ryoo. Too many frames, not all useful: Efficient strategies for long-form video qa. *arXiv preprint arXiv:2406.09396*, 2024. 2
- [47] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 1
- [48] Ye Qinghao, Xu Guohai, Yan Ming, Xu Haiyang, Qian Qi, Zhang Ji, and Huang Fei. Hitea: Hierarchical temporal-aware video-language pre-training. *arXiv preprint arXiv:2212.14546*, 2022. 2
- [49] 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*. PMLR, 2021. 1, 6
- [50] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*, 2019. 3, 6
- [51] Ivan Rodin, Antonino Furnari, Kyle Min, Subarna Tripathi, and Giovanni Maria Farinella. Action scene graphs for long-form understanding of egocentric videos. In *CVPR*, 2024. 1
- [52] Tim Salzmann, Markus Ryll, Alex Bewley, and Matthias Minderer. Scene-graph vit: End-to-end open-vocabulary visual relationship detection. In *ECCV*. Springer, 2025. 3
- [53] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *ICMR*, 2019. 3, 5
- [54] Gunnar A Sigurdsson, GÜl Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 3, 5
- [55] Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *NAACL*, 2021. 6
- [56] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *ACM*, 2016. 3
- [57] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 7
- [58] Aisha Urooj, Hilde Kuehne, Bo Wu, Kim Chheu, Walid Boushelham, Chuang Gan, Niels Lobo, and Mubarak Shah. Learning situation hyper-graphs for video question answering. In *CVPR*, 2023. 1, 2, 3, 6
- [59] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric

- Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17, 2020. 1
- [60] Han Wang, Yanjie Wang, Yongjie Ye, Yuxiang Nie, and Can Huang. Elysium: Exploring object-level perception in videos via mllm. *ECCV*, 2024. 1
- [61] 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 *CVPR*, 2023. 6
- [62] Shijie Wang, Qi Zhao, Minh Quan Do, Nakul Agarwal, Kwon-joon Lee, and Chen Sun. Vamos: Versatile action models for video understanding. *arXiv preprint arXiv:2311.13627*, 2023. 2, 5, 6, 7
- [63] Xijun Wang, Junbang Liang, Chun-Kai Wang, Kenan Deng, Yu Michael Lou, Ming Lin, and Shan Yang. Vila: Efficient video-language alignment for video question answering. *ECCV*, 2024. 2, 6, 7
- [64] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyuan Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 2, 6, 7
- [65] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *ECCV*, 2024. 2
- [66] Zhenhailong Wang, Ansel Blume, Sha Li, Genglin Liu, Jaemin Cho, Zineng Tang, Mohit Bansal, and Heng Ji. Paxion: Patching action knowledge in video-language foundation models. *NeurIPS*, 36, 2023. 1
- [67] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *NeurIPS*, 2021. 1, 2, 5
- [68] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. *CVPR*, 2024. 2
- [69] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021. 1, 2, 5
- [70] Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. Video graph transformer for video question answering. In *ECCV*. Springer, 2022. 2, 7
- [71] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueling Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM MM*, 2017. 2
- [72] Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. Slowfast-llava: A strong training-free baseline for video large language models. *arXiv preprint arXiv:2407.15841*, 2024. 2, 5, 6
- [73] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *ICLR*, 2025. 2, 6
- [74] Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. Hitea: Hierarchical temporal-aware video-language pre-training. In *ICCV*, 2023. 6
- [75] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *NeurIPS*, 36, 2024. 2, 6, 7
- [76] Zhou Yu, Lixiang Zheng, Zhou Zhao, Fei Wu, Jianping Fan, Kui Ren, and Jun Yu. Anetqa: A large-scale benchmark for fine-grained compositional reasoning over untrimmed videos. In *CVPR*, 2023. 1, 3
- [77] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *ICLR*, 2022. 1
- [78] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *CVPR*, 2022. 2, 7
- [79] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *EMNLP*, 2023. 2, 6
- [80] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *ICLR*, 2024. 5, 7

REVEAL: Relation-based Video Representation Learning for Video-Question-Answering

Supplementary Material

Acronyms

BLIP Bootstrapping Language-Image Pre-training

CLIP Contrastive Language-Image Pre-training

CLS Classification

FFN Feed Forward Network

LLM Large Language Model

MIL-NCE Multiple Instance Learning and Noise Contrastive Estimation

MM-NCE Many-to-Many Noise Contrastive Estimation

MSE Mean Squared Error

REVEAL RElation-based Video rEpresentAtion Learning

VideoQA Video-Question-Answering

VLM Video Language Models

6. Pretraining Details

6.1. Dataloading

Our data preprocessing and loading pipeline relies on WebDataset [1]. Precomputed slow-fast CLIP features are stored as TAR files containing PyTorch [47] tensors. We use WebDataset’s built-in shuffling mechanism with a buffer size of 5000 samples and an initial buffer of 1000 samples to ensure proper randomization.

6.2. Model Implementation

The pretraining model architecture consists of dual-pathway transformers processing slow and fast video features. We extract CLIP’s patch features from the penultimate layer for the slow pathway. Each pathway includes a projection layer that maps 1024-dimensional input features to a hidden dimension 768, followed by learnable positional encodings. The fast pathway processes CLS tokens features from 16 frames, while the slow pathway handles patch features from 4 frames. Both pathways utilize identical but separate transformer encoders, each comprising two encoder layers with 8-head self-attention (hidden size 768, FFN dimension 4×768). The model employs fixed positional encodings using sinusoidal functions. We implement separate embedding modules for relationship modeling, generating 8 learnable query embeddings for each pathway. The decoder architecture comprises

12 transformer decoder layers per pathway, each with 8-head cross-attention mechanisms and GELU activation functions. The decoder outputs are processed through an MLP with architecture $768 \rightarrow 4 \times 768 \rightarrow 1024$, where 1024 is the ground truth embedding dimension. The implementation includes careful initialization strategies: orthogonal initialization for query embeddings, normal initialization (mean=0, std=0.02) for linear layers, and zero for biases. All normalization layers use LayerNorm.

6.3. Pretraining

Our pretraining implementation utilizes distributed training using PyTorch’s DistributedDataParallel (DDP). The learning rate follows a cosine schedule with a linear warmup, starting from an initial learning rate of $lr = 0.00005$ with a 20% warmup period over total steps, decaying to $0.05 \times lr$ at completion. Training proceeds for 5 epochs with gradient accumulation every 4 steps and gradient clipping at 1.0. We implement a bidirectional contrastive loss adapted to our multi-prediction setting following open-clip implementation [8]. We use the Hungarian matching implementation from Scipy [59] to match predictions with ground truth. The model employs two separate prediction heads for slow and fast pathways, each producing embeddings of dimension 1024. We initialize the logit scale as $\log(1/0.07)$. We use the AdamW optimizer with a weight decay of 0.1. Training metrics are logged using Neptune.ai [45], including gradient norms, learning rates, and various losses.

7. Finetuning Details

Our implementation leverages Llama-VQA implementation [26]. Llama3 8B is the base language model, enhanced with REVEAL for video processing. We fine-tune using adapter layers while keeping the base Llama model frozen. Specifically, we use 32 adapter layers, with a length of tokens corresponding to the number of video relationships input to the LLM. The model extracts 16 relation queries per temporal segment, which are then linearly projected to match Llama’s hidden dimension (4096). The training process uses AdamW optimizer with a base learning rate scaled by batch size (effective $lr = base_lr \times batch_size/256$), with a linear warmup over 2 epochs and cosine decay. The training is done for five epochs. We use slow-fast features with a dimension of 1024 for video features, which are processed through REVEAL before being integrated with the language model. For datasets requiring subtitles (TVQA and VLEP), we integrate them into the input sequence before the ques-

tion. All video frame features are pre-extracted and stored. In table 8, we provide the hyperparameters per dataset.

Hyperparameter	STAR	NextQA	Intent-QA	TVQA	VLEP
Base Learning Rate	0.06	0.06	0.08	0.07	0.07
Batch Size	4	8	4	1	4
Weight Decay	0.14	0.1	0.14	0.02	0.12
Temporal Resolution	8	2	2	1	1
Gradient Accum.	8	4	4	4	2
Bias	3	3	3.5	3	3
QAV loss	✓	✓	✓	✓	✓
VAQ loss	✓	✓	✓	✓	✗
Max Sequence Length	256	192	256	714	384

Table 8. Dataset-specific hyperparameters used in our experiments. Values were determined through empirical validation.

8. Full Ablation Tables

Table 10 provides full per-category results for the temporal resolution, the relationship encoder initialization, and the number of relationships input to the LLM. The optimal temporal resolution, as expected intuitively, depends on the dataset. We also observe that the model with a relationship encoder initialized from a sentence embedder improves the performance of every question category evaluated. Finally, the more relationship vectors we input to the LLM, the better the results are, even though we get competitive results from a single relationship vector per temporal segment.

9. Qualitative Analysis of the Performance on VideoQA

9.1. Successful Cases

We present two successful examples from the STAR dataset where our model correctly answers the questions (Figure 4). In both cases, we visualize the alignment between the extracted relationship triplets and video segments (*i.e.*, the maximum similarity scores between the decoded queries and the encoded relationships) to demonstrate how REVEAL processes temporal information. In the first example, given the question "What is the person doing while eating a sandwich?", we extract the relationship triplet "Subject: person, Predicate: eating, Object: sandwich". The video is divided into 8 equal segments, and we observe strong alignment between this relationship and all segments, confirming the continuous eating action. The correct answer, "took blanket", shows increased alignment specifically during the relevant temporal window, while alternative choices exhibit lower alignment scores as these actions are absent in the video. In the second example, for the question "What happened before the person opened the door?", we observe that the question's relationship becomes well-aligned with the video during the final two segments, corresponding to the door-opening action. The correct answer, "sat at the table", shows stronger

alignment during the first six segments, maintaining higher scores than incorrect choices.

9.2. Failure Cases

We also analyze two failure cases from the STAR dataset (Figure 5) to understand the model's limitations. The first case involves question ambiguity: given "What is the person doing after touching the box?", the model predicts "put down the box" while the ground truth is "closed the box". The alignment plots show that both relationships are well-matched with the video content, suggesting that both answers could be valid interpretations of the observed action sequence. The second example ("What is the person doing after opening the closet?") demonstrates an object recognition challenge. While the correct answer involves taking a box, the model incorrectly predicts "take clothes". The alignment reveals the model's difficulty in recognizing the box, and the prediction may be influenced by Llama3's knowledge about items typically retrieved from closets.

10. Prompt Engineering and Relationship Extraction

10.1. Prompt Template

Figure 7 shows the complete prompt template used with Mistral-7B for relationship extraction. We leverage in-context learning by providing multiple examples of caption-relationship pairs before requesting the model to extract relationships from new captions. Each example demonstrates how to decompose a caption into subject-predicate-object triplets. The prompt includes diverse examples covering different types of actions, objects, and temporal relationships to encourage comprehensive extraction. This approach helps the model understand the expected format and granularity of the extracted relationships.

10.2. Relationship Extraction Examples

Table 9 presents examples of relationships extracted from Webvid-2M captions, with corresponding video frames shown in Figure 6. The extraction results demonstrate several key properties of our approach. The LLM generates a focused set of core relationships for concise captions. In contrast, complex or longer captions yield more detailed relationship sets. The extracted relationships, while accurate, are not exhaustive - they do not cover every possible relationship that could be inferred from the video content. This non-exhaustive nature of the extracted relationships validates our design choice not to penalize missing relationships during training to let the model freely infer relevant relationship vectors from videos.



Figure 4. Successful examples from STAR dataset demonstrating REVEAL’s relationship alignment capabilities. Top: The model correctly identifies concurrent actions (eating sandwich while taking blanket). Bottom: The model successfully captures temporal ordering of actions (sitting at table before opening door). Alignment scores between extracted relationships and video segments are visualized, showing stronger alignment during relevant temporal windows.

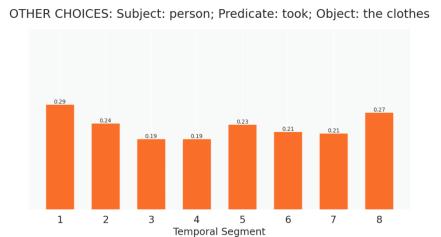
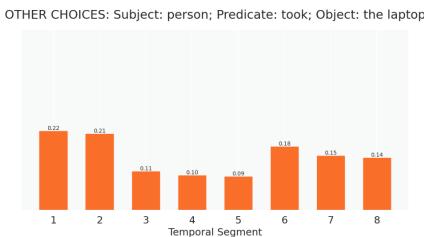
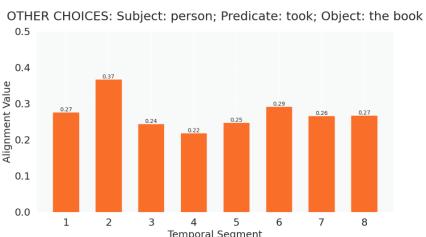
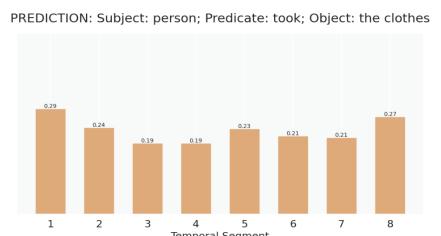
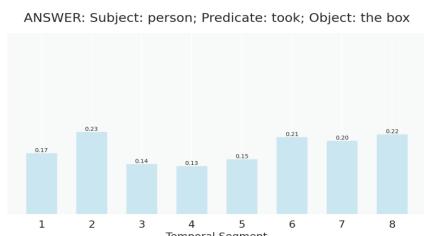
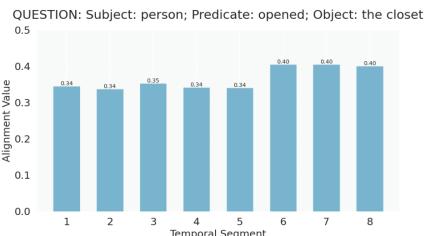
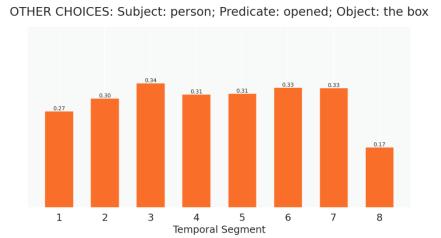
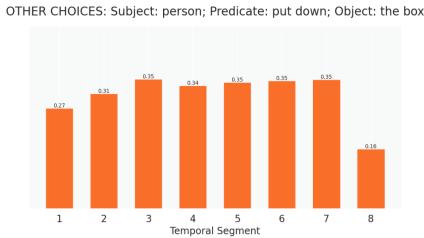
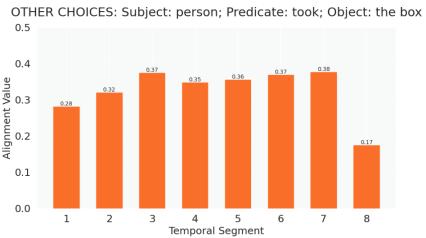
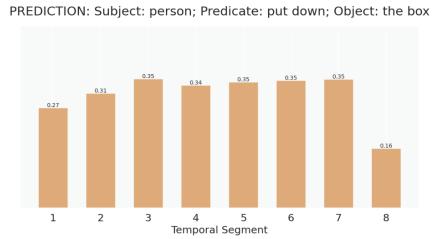
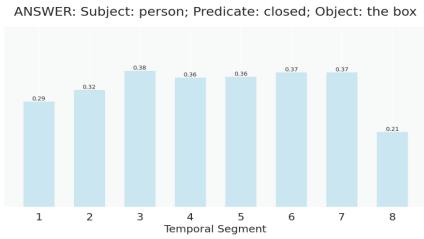
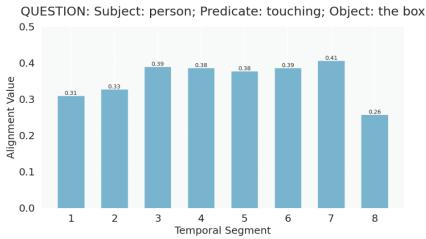


Figure 5. Failure cases from STAR dataset highlighting REVEAL’s limitations. Top: Question ambiguity leads to multiple valid interpretations of the same action sequence. Bottom: Object recognition challenge where the model defaults to common-sense assumptions about closet contents rather than recognizing the specific object (small box).

Caption	Relationships
Roses in blossom slow motion cinematic video	<ul style="list-style-type: none"> • Subject: roses , Predicate: in blossom, Object: none • Subject: roses , Predicate: appearing in, Object: cinematic video • Subject: cinematic video, Predicate: having, Object: slow motion
Male showing yellow particles inside the body showing the cardiovascular system, lungs, heart, liver, stomach and intestines with radar graphic below and shining light from the top left corner	<ul style="list-style-type: none"> • Subject: male, Predicate: showing, Object: yellow particles • Subject: male, Predicate: showing, Object: body • Subject: body, Predicate: showing, Object: cardiovascular system • Subject: body, Predicate: showing, Object: lungs • Subject: body, Predicate: showing, Object: heart • Subject: body, Predicate: showing, Object: liver • Subject: body, Predicate: showing, Object: stomach • Subject: body, Predicate: showing, Object: intestines • Subject: radar graphic, Predicate: below, Object: male
Iguana on a tree hd	<ul style="list-style-type: none"> • Subject: iguana, Predicate: on, Object: tree
Turtle and tortoise on stone decoration design in pond of garden japanese style in naritasan plum garden of narita public park at chiba prefecture in tokyo, japan	<ul style="list-style-type: none"> • Subject: turtle and tortoise, Predicate: on, Object: stone decoration • Subject: turtle and tortoise, Predicate: in, Object: pond • Subject: pond, Predicate: of, Object: garden • Subject: garden, Predicate: japanese style, Object: None • Subject: garden, Predicate: in, Object: Narita public park • Subject: Narita public park, Predicate: at, Object: Chiba prefecture
Polishing of wooden plank using a rasp	<ul style="list-style-type: none"> • Subject: person, Predicate: polishing, Object: wooden plank
Athletic woman in sportswear holding feet on box and doing evaluated reverse plank with leg raise while training at outdoor fitness court	<ul style="list-style-type: none"> • Subject: athletic woman, Predicate: holding, Object: feet • Subject: athletic woman, Predicate: doing, Object: reverse plank • Subject: athletic woman, Predicate: raising, Object: leg • Subject: box, Predicate: under, Object: feet • Subject: outdoor fitness court, Predicate: at, Object: training
Canada goose family walking with the amazing view of mount cook (aoraki)	<ul style="list-style-type: none"> • Subject: Canada goose family, Predicate: walking • Subject: Canada goose family, Predicate: with, Object: amazing view • Subject: amazing view, Predicate: of, Object: mount cook (aoraki)
Extreme close up image with chess game pieces moved on the board by player hand	<ul style="list-style-type: none"> • Subject: player, Predicate: moving, Object: chess game pieces • Subject: player, Predicate: taking, Object: chess game pieces • Subject: chess game pieces, Predicate: on, Object: board • Subject: image, Predicate: close up • Subject: image, Predicate: containing, Object: chess game pieces and player hand • Subject: image, Predicate: having, Object: extreme close up perspective
Aerial view of a beautiful beach with turquoise water and waves crashing on the shore	<ul style="list-style-type: none"> • Subject: view, Predicate: aerial, Object: beach • Subject: beach, Predicate: is, Object: beautiful • Subject: water, Predicate: is, Object: turquoise • Subject: waves, Predicate: crashing on, Object: shore

Table 9. Video Captions From Webvid-2M and Their Extracted Relationships

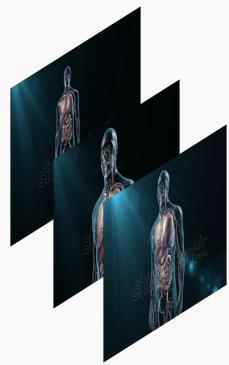
10.3. Relationship Extraction Pipeline for Charades and VidOR

The VidOR and Charades datasets provide temporal annotations of relationships between objects in videos. Each relationship is annotated by a subject-predicate-object triplet and its temporal extent (start and end frames). To process these relationships into meaningful clips, we first collect all temporal ranges (t_{start}, t_{end}) for each video. We then

employ a dynamic grouping algorithm that identifies natural breaks in the temporal annotations by analyzing the gaps between consecutive relationships. Specifically, we calculate the gap sizes between temporally adjacent relationships and use the 75th percentile of these gaps as a threshold to determine significant temporal breaks. This approach naturally segments the video into clips containing temporally coherent relationships. For each resulting clip, we aggregate



Roses in blossom slow motion cinematic video



Male showing yellow particles inside the body showing the cardiovascular system, lungs, heart, liver, stomach and intestines with radar graphic below and shining light from the top left corner



Iguana on a tree hd



Turtle and tortoise on stone decoration design in pond of garden Japanese style in naritasan plum garden of narita public park at chiba prefecture in tokyo, japan



Polishing of wooden plank using a rasp



Athletic woman in sportswear holding feet on box and doing evaluated reverse plank with leg raise while training at outdoor fitness court



Canada goose family walking with the amazing view of mount cook (aoraki)



Extreme close up image with chess game pieces moved on the board by player hand



Aerial view of a beautiful beach with turquoise water and waves crashing on the shore

Figure 6. Sample videos from WebVid-2M

all relationships whose temporal extent overlaps with the clip's timeframe, creating a set of relationships that describe

the scene dynamics within that temporal window.

Ablation	STAR					NeXT-QA				Intent-QA			
	In	Seq	Pre	Feas	All	C	T	D	All	CW	CH	TP & TN	All
a) Temp. Res.:													
1	54.9	62.3	64.1	65.9	61.8	74.2	68.8	77.0	73.4	74.3	61.0	55.0	70.7
2	57.6	65.1	69.4	68.4	65.1	74.0	70.0	77.9	73.3	74.6	75.5	65.6	71.8
4	59.2	68.0	70.7	69.0	66.7	73.5	69.2	76.7	72.6	74.3	74.7	60.3	71.1
8	61.4	69.3	75.0	72.0	69.4	73.7	69.6	75.4	72.7	72.2	75.0	60.6	70.8
b) Rel. Init.:													
Random	59.7	68.2	72.9	72.5	68.3	72.4	67.1	74.4	71.0	70.7	73.4	58.7	69.2
RoBERTa-large	60.4	67.9	74.0	69.6	68.0	73.8	68.4	74.9	72.2	73.7	75.9	60.2	71.0
CLIP text encoder	59.4	69.0	73.1	72.5	68.5	74.0	68.2	74.9	72.3	72.2	75.9	61.0	71.4
Sentence embedder	61.4	69.3	75.0	72.0	69.4	74.0	70.0	77.9	73.3	74.6	75.5	65.6	71.8
c) #Rels:													
1	57.8	66.9	68.4	67.8	65.3	73.8	68.0	77.2	72.4	72.5	74.1	60.8	70.5
2	62.1	68.4	74.0	68.4	68.3	74.1	68.9	77.1	72.9	74.9	74.3	60.2	70.7
4	60.4	67.7	73.7	70.2	68.0	75.1	68.3	75.9	73.1	74.9	74.4	61.8	71.3
8	61.4	69.3	75.0	72.0	69.4	74.0	70.0	77.9	73.3	74.6	75.5	65.6	71.8

Table 10. Full per category results for a) Temporal resolution; b) Relationship encoder initialization and; c) number of relationships vectors input to the LLM.

Relationships Extraction Prompt

```
[INST] You are a software to extract relationships from sentences.
Extract explicit and factual relationships between objects in the last sentence.
Use the same formatting as below. No other text.
One instance per subject, object, and predicate. Be exhaustive.

Sentence: 'A video of a person on the side of a table holding food.'
subject: person, predicate: on the side of, object: table
subject: person, predicate: holding, object: food

Sentence: 'A kid touching the table while sitting on a chair.'
subject: kid, predicate: touching, object: table
subject: kid, predicate: sitting on, object: chair

Sentence: 'A man putting on shoes and clothes.
Behind him two trees next to each other.'
subject: man, predicate: holding, object: shoe
subject: man, predicate: holding, object: clothes
subject: two trees, predicate: behind, object: him
subject: tree, predicate: next to, object: tree

Sentence: 'Woman sets table with plates, silverware, glasses,
before placing oatmeal pot and juice pitcher in center. Calls family.'
subject: woman, predicate: set, object: table
subject: woman, predicate: set, object: plates
subject: woman, predicate: set, object: silverware
subject: woman, predicate: set, object: glasses
subject: woman, predicate: placing, object: oatmeal pot
subject: woman, predicate: placing, object: juice pitcher
subject: oatmeal pot, predicate: in center of, object: table
subject: juice pitcher, predicate: in center of, object: table
subject: woman, predicate: call, object: family

Sentence: 'Children playing on swings and slide. Couple sits on bench,
holding hands.'
subject: children, predicate: playing on, object: swings
subject: couple, predicate: sit on, object: bench
subject: couple, predicate: holding, object: hands [/INST]
Sentence: {sentence}
```

Figure 7. Prompt for Extracting Relationships from Sentences