Causal and Temporal Inference in Visual Question Generation by Utilizing Pre-trained Models

Zhanghao Hu and **Frank Keller**
School of Informatics, University of Edinburgh, UK
huzh666295@gmail.com, keller@inf.ed.ac.uk

Abstract

Visual Question Generation is a task at the crossroads of visual and language learning, impacting broad domains like education, medicine, and social media. While existing pre-trained models excel in fact-based queries with image pairs, they fall short of capturing human-like inference, particularly in **understanding causal and temporal relationships** within videos. Additionally, the computational demands of prevalent pre-training methods pose challenges. In response, our study introduces a framework that leverages vision-text matching pre-trained models to guide language models in recognizing event-entity relationships within videos and generating inferential questions. Demonstrating efficacy on the NExT-QA dataset, which is designed for causal and temporal inference in visual question answering, our method successfully guides pre-trained language models in recognizing video content. We present methodologies for abstracting causal and temporal relationships between events and entities, pointing out the importance of consistent relationships among input frames during training and inference phases and suggesting an avenue for future exploration[1].

1 Introduction

Visual Question Generation (VQG) is an emerging task of multi-modal learning, integrating vision and language. Since its inception (Lin and Parikh, 2016), VQG has influenced diverse domains like education (Zhao et al., 2022), social media (Yeh et al., 2022), and human-computer interaction (Lee et al., 2018). Existing datasets primarily cater to factoid question answering, extracting direct answers from visual content (Yeh et al., 2022). However, factoid question answering lacks inherent depth in human thinking, exemplified by the disparity between a fact-based query like "Was anyone injured in the crash?" and a more insightful, causal question "Why do these drivers have accidents in the middle of intersections?" or a temporal question "What will the police do after the crash?"

This research addresses a critical gap in the VQG landscape: *the absence of studies exploring inference aligned with human thinking.* Moreover, unlike singular images, videos offer richer details of relationships between events and entities, prompting our focus on two fundamental types of reasoning—*causal inference and temporal inference*. Through this approach, we aim to introduce a new challenge of inferential question generation originating from videos and auxiliary text and advance the field of VQG.

Meanwhile, despite advancements in VQG, the computational demands (Radford et al., 2021) of pre-training models, particularly visual transformers (Dosovitskiy et al., 2020), pose challenges. Our work distinguishes itself by harnessing pre-trained vision-to-text matching models instead of embarking on resource-intensive model training from scratch. Inspired by prior successes (Mokady et al., 2021), our approach expedites question generation by leveraging the knowledge embedded in existing models, thereby enhancing the quality and efficiency of the process.

The contributions of this paper are as follows:

1. As far as we know, we are the first to explore the task of causal and temporal video question generation. We propose a framework (figure 1) and establish a baseline step by step by comparing video encoders, language model sizes, and stage fine-tuning strategies. Additionally, we propose an evaluation metric to enhance VQG grounding assessment.

2. Experiments on the NExT-QA dataset display the efficacy of our methods in combining vision and language. *We highlight the importance of consistent frame relationships during training and inference for deriving event-*

[1]The code is available at this address.

Proceedings of the 3rd Workshop on Advances in Language and Vision Research, pages 138–154

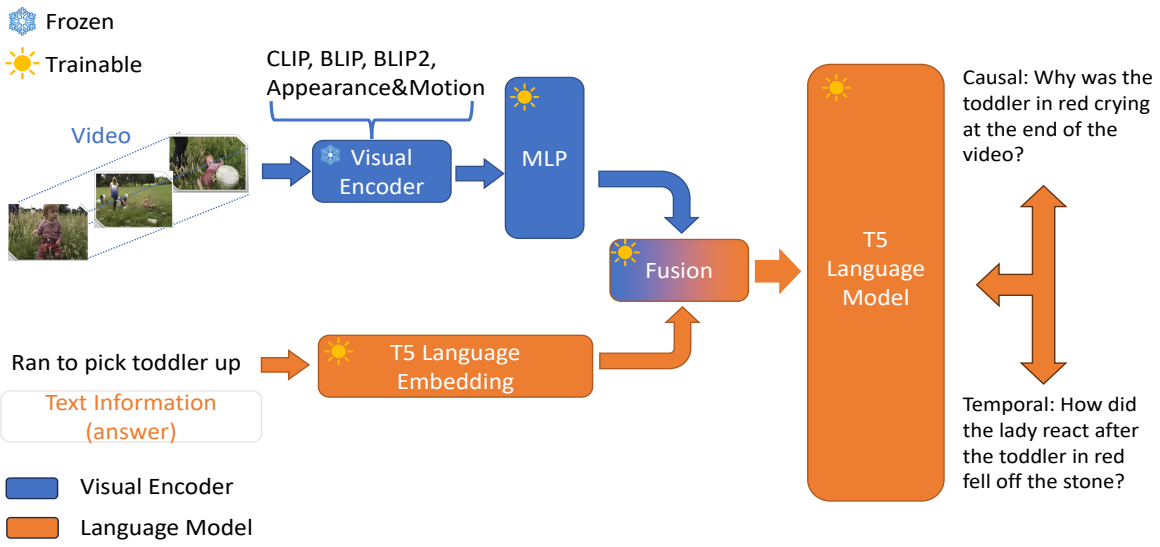


Figure 1: The overall framework for visual question generation. It comprises four essential components: a visual encoder, an auxiliary text encoder (T5), multi-modal interaction, and an output question decoder (T5). Videos and auxiliary text are respectively encoded into embeddings and be concatenated through multilayer perception (MLP) layers. Temporal and causal questions will be generated by the question decoder.

entity relationships within videos. This research suggests the direction of enhancing frame-based consistency in causal and temporal video inference for future work.

2 Background and Related Work

Visual Question Generation: The field of VQG has seen notable progress since its introduction (Mostafazadeh et al., 2016). Existing research has extensively explored single-image VQG (Vedd et al., 2021; Krishna et al., 2019), while multiple-image VQG (Chan et al., 2022) and video VQG (Khurana and Deshpande, 2021), which present promising avenues for inferring causality and temporal relationships between visual elements, remain unexplored. To the best of our knowledge, no prior research has specifically focused on the challenges of generating questions that involve causal and temporal inference in VQG tasks. This represents a critical research gap, as inferential questions have the potential to unlock deeper insights of visual content, going beyond mere factual queries.

Multi-modal Generative Task with Pre-trained Models: Existing research in visual question generation adopts large pre-trained models for tasks like image captioning (Li et al., 2022), visual question answering (Khan et al., 2023), and visual grounding (Peng et al., 2023), showcasing impressive results but facing high computational costs (Dosovitskiy et al., 2020). An alternative, leveraging vision-text matching pre-trained models like CLIP (Radford et al., 2021), BLIP (Li et al., 2022), and BLIP2 (Li et al., 2023), efficiently bridges vision and language domains. Despite success in various generative tasks, no prior research explores these models for vision-based question generation, particularly those involving causal and temporal inference. This research aims to utilize various vision-text matching pre-trained models in capturing causal and temporal relationships.

3 Methods

The overall framework for VQG is displayed in Figure 1. This section introduces our training strategies and inferential relationship abstraction methods.

3.1 Multi-modal Fusion

Visual information and textual context are often complementary in nature. The visual content provides rich details and cues that are not present in the text, and vice versa. The core issues are *how to unify the multi-modal embedding space between vision and language*, and how to *effectively guide the language model in recognizing visual information and generating temporal and causal questions*.

Concatenate Vision and Language: Inspired by one of the latest methods (Liu et al., 2023b,a),

we propose a direct but powerful technique to connect vision and language spaces. Specifically, given auxiliary text input words which are "Text Information" shown in Figure 1 $w_V^1, w_V^2, ..., w_V^p$, for a video V, we process them by language models and get a series of word embeddings $t_V^1, t_V^2, ..., t_V^i$. Given a video V, we first divide the video V as separate frames $x_V^1, x_V^2, ..., x_V^m$. Next, after processing the frames by visual encoders, we employ a light mapping network (multilayer perceptron), denoted by F, to map the visual embedding to k embedding vectors (we set the k as 5 in our experiments):

$$p_V^1, p_V^2, ...p_V^k = F(visual\_encoder(x_V^1, ..., x_V^m)). \quad (1)$$

where each vector p_V^k has the same dimension as the word embedding of language models. We then concatenate the obtained visual embedding to the auxiliary input text embeddings:

$$Z_V = p_V^1, ..., p_V^k, t_V^1, ..., t_V^i. \quad (2)$$

During fine-tuning, we feed the language models with the prefix-text concatenation $\{Z_i\}_{i=1}^N$, where N is the number of videos. Our training objective is to predict the temporal and causal question tokens conditioned on the prefix in an auto-regressive fashion. To this purpose, we train the mapping component F using the simple, yet effective, cross-entropy loss:

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{j=1}^{\ell} \log p_\theta(q_j^i | Z_V, q_1^i, ...q_{j-1}^i), \quad (3)$$

where ℓ is the length of the predicted questions, p_θ is the probability of ground-truth tokens,.

Two Stage Fine Tuning: Inspired by prior research (Liu et al., 2023b,a), a two-stage fine-tuning methodology is introduced to tackle the challenge of multi-modal fusion in visual question generation by effectively aligning visual and textual information. In the first stage, we prioritize feature alignment fine-tuning, aligning the visual encoder with the language model through a parameter mapping network F. This ensures alignment between video features and language model word embeddings, streamlining visual tokenizers. In the second stage, a fine-tuning end-to-end strategy takes place after the convergence of the first stage. Visual encoder weights are frozen, and both pre-trained weights of the projection layer and the language model are updated. This two-stage process, acting on the "Fusion model" shown in Figure 1, optimizes the language model's performance.

3.2 Causal and Temporal Inference Abstraction Methods

This section introduces two methods which aim to enhance the abstraction of causal and temporal inference from events and entities within a video.

Vision Projection Matrix Choice: An intuitively straightforward approach is taken by creating distinct MLP layers for individual frames similar to equation 1 (In this experiment we set the number of the MLP layers as 16), aiming to capture nuanced characteristics. Each frame's embeddings are projected onto a linguistic embedding using an additional MLP with a prefix length of 5.

Contradictory Frame Comparison aims to abstract causal and temporal relationships in a video by exploiting differences between consecutive frames. Two strategies are employed using the CLIP vision encoders. (1) *Global Frame Comparison*: 16 frames at uniform intervals are transformed into vision embeddings through the CLIP encoder. Pairs of frames with the lowest cosine similarity represent the most contradictory frames, projected onto the language embedding through an MLP layer. (2) *Local Frame Comparison*: Once again, we select pairs of frames and calculate their cosine similarity. But during training, firstly the CLIP model is invoked to determine the most relevant frame in relation to the given question and answer since at training time we have all relevant inputs. Then, we select the rest frame which displays the lowest cosine similarity with the contextually chosen frame. Again, an MLP layer projects the selected frame pair onto the language embedding.

4 Experiment

4.1 Data and Evaluation

Existing video question-answering datasets primarily address factoid questions with direct visual answers (Xu et al., 2016; Jang et al., 2017) but lack inference questions. To fill this gap and integrate causal and temporal inference, this study opts for the **NExT-QA** dataset (Xiao et al., 2021), which is designed for inferential visual-question-answering, offering about 52K diverse questions (48% causal, 29% temporal, 23% descriptive).

The assessment of visual question generation (VQG) systems traditionally relies on language metrics like BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), and CIDEr (Vedantam et al., 2015), designed for machine translation, lacking inference

evaluation. To address this gap, our study introduces new metrics—precision, recall, and F1-score grounding—examining word overlap between predicted and ground-truth questions. The grounding metrics consider matching overlaps of content-bearing words and exclude irrelevant words[2]. We define the formula of the grounding metrics:

$$\begin{aligned} PG &= N_{\text{matching overlap}} / N_{\text{predicted question tokens}} \\ RG &= N_{\text{matching overlap}} / N_{\text{ground truth question tokens}} \\ FG &= \frac{2 * PG * RG}{PG + RG}, \end{aligned} \tag{4}$$

Where PG means Precision Grounding, RG means Recall Grounding, FG means F1 score Grounding, $N_{\text{matching overlap}}$ counts matching overlaps between predicted and ground truth questions. $N_{\text{predicted question tokens}}$ and $N_{\text{ground truth question tokens}}$ represent the respective token counts.

4.2 Experiment Setup

Baseline Models: In establishing baseline models for a fair comparison on the NExT-QA datasets, we employ the **Heterogeneous Graph Attention (HGA) model** (Jiang and Han, 2020) and **a pre-trained language model** with text-only input: (1) The HGA model utilizes 3D motion and 2D appearance vectors, abstracted from ResNet (He et al., 2016) and ResNeXt-101 (Xie et al., 2017).(2) The pre-trained language model T5 (Radford et al., 2021) is explored with text-only input as a baseline, to assess its ability to recognize visual content in videos in the following experiments.

Video Encoder: To enhance visual question generation for temporal and causal inference, traditional 2D and 3D convolutional networks face limitations in generative tasks. Leveraging pre-trained vision-text matching models like **CLIP (Radford et al., 2021), BLIP (Li et al., 2022) and BLIP2 (Li et al., 2023)**, we conduct a comprehensive performance comparison against convolutional networks.

Language Model Size Selection: To explore the impact of language model size on recognizing relationships in videos, we employ **T5 Small** and **T5 Large**. In addition, we adopt two tuning strategies. "One Stage" in Table 4 and Table 5 means we directly train the mapping network F in section 3.1 from scratch and "Two Stage" represents the fine-tuning strategy explained in section 3.1.

[2]We exclude the words of POS types "CC", "DT", "IN", "TO" and "UH" in our experiments.

4.3 Experiment Results

4.3.1 Baseline

We evaluated our baseline models with results summarized in Table 1. The HGA model, incorporating video and text input, achieves the highest grounding score but exhibits lower question quality due to stop-word repetition and shorter length generation (Figure 2). Although BLEU has a brevity penalty and METEOR and ROUGEL consider the recall evaluation metrics, with higher precision, the evaluation performance of the HGA model still gets close to that of the T5 model. In addition, as shown in Table 2, since our grounding metric ignores stop-words and considers only relevant words to the vision content such as nouns and verbs, precision will have an advantage in the evaluation compared to recall, thus the HGA model achieves a significant improvement compared to the T5 model. However, HGA has comparatively lower recall than those of T5 in causal and temporal question generation (Table 2). In conclusion, HGA exhibits higher precision and F1-score in the grounding metric but lower performance in BLEU, METEOR, CIDEr, and recall in the grounding metric of causal and temporal questions. *This leads us to choose T5 as the foundation for subsequent experiments.*

Model	B	RL	M	C	Grounding
HGA	0.1248	**0.4128**	0.3101	0.8271	**0.3248**
T5 Small Text Only	**0.1269**	0.3857	**0.3276**	**0.8480**	0.2957
T5 Large Text Only	0.1239	0.3851	0.3237	0.8353	0.2987

Table 1: Baseline Model Evaluation Performance. B is BLEU, RL is ROUGEL, M is METEOR, C is CIDEr, and Grounding is the F1-score grounding metric.

Model	C G-Pre	C G-Re	C G-F1	T G-Pre	T G-Re	T G-F1
HGA	**0.3378**	0.2357	**0.2776**	**0.4126**	0.2763	**0.3310**
T5 Small Text Only	0.2527	0.2541	0.2534	0.3096	**0.2943**	0.3018
T5 Large Text Only	0.2736	**0.2650**	0.2692	0.2998	0.2786	0.2888

Table 2: Baseline Model Grounding Performance in Causal and Temporal Inference. C G represents the grounding metric of causal questions. T G represents the grounding metric of temporal questions. "Pre" represents precision. "Re" represents recall. "F1" represents the F1 score.

4.3.2 Multi-modal Concatenation

Video Encoder Comparison: We assess the performance of different vision video encoders, and

Video:

Ground Truth Question:
1:how did the girl react when she saw that the ball was running at the beginning?
2: why was the girl on the floor in the middle of the video?
3: why does the dog chase the ball?
4: why did the girl run down the slope?
5: why does the child run after the ball?
6: what did the dog do after the girl approached the adult and stood beside him?
7: where is this video taken?
8: what did the girl do after she fell on the ground?
9: what did the girl do after she finished playing with the ball at the end of the video?
10: what did the girl do to the dog when the dog stood beside the man?
11: what did the girl do after the dog took the ball away?
12: why does the child run down the slope after the ball rolls away from him?

HGA: ***42 matching overlap***
"1": "what did the boy do after the girl ran away",
"2": "why did the girl in after the girl in the",
"3": "why did the dog run towards the ball",
"4": "why did the boy run to the ball after the ball",
"5": "why did the boy run down the ball",
"6": "what did the boy do after the ball ball",
"7": "where is this video taken",
"8": "what did the dog do after the the ball",
"9": "what did the girl do after the the ball",
"10": "what did the boy do after the dog ran away",
"11": "what did the girl do after the dog ran away",
"12": "what did the dog do after the ball ball"

T5-small text only: ***27 matching overlap***
"1": "what did the boy do after he walked away from the ball",
"2": "why did the girl in pink hold onto the girl in pink when she is squatting down",
"3": "why did the baby put his hand on the toy in the middle of the video",
"4": "why did the man in black bend down at the start of the video",
"5": "why did the man in black bend down at the start of the video",
"6": "what does the man in black do after the man in black starts talking",
"7": "where is this video taken",
"8": "what did the boy do after he walked to the other side of the room",
"9": "what does the girl do after the girl in pink starts dancing",
"10": "what does the man do after the dog starts running",
"11": "what does the dog do after the dog starts running",
"12": "what does the man in black do after the man in black starts playing the drums"

Figure 2: Baseline Performance. Yellow markup shows the matching overlap compared with the ground truth questions. Red markup shows the repetitive words.

the results are summarized in Table 3. CLIP and BLIP2 stand out, with CLIP excelling in ROUGEL, and Grounding metrics, showcasing good visual content recognition. In contrast, BLIP2 performs well in BLEU, METEOR, and CIDEr, generating detailed questions. Despite BLIP2's detailed questions, CLIP's higher matching overlap with ground truth and its balanced performance led to the *selection of CLIP as the video encoder for subsequent experiments* (Figure 3).

Model	B	RL	M	C	Grounding
None Text Only	0.1269	0.3857	0.3276	0.8480	0.2957
App&Mot	0.1348	0.3958	0.3353	0.8816	0.3092
CLIP	0.1564	**0.4216**	0.3594	1.0366	**0.3505**
BLIP	0.1562	0.4179	0.3584	1.0205	0.3425
BLIP2	**0.1583**	0.4210	**0.3599**	**1.0488**	0.3455

Table 3: Visual encoders performance with T5-small following Section 3.1 fusion method. App&Mot means 2D appearance vectors and 3D motion vectors abstracted from convolution networks. B is BLEU, RL is ROUGEL, M is METEOR, C is CIDEr, and Grounding is the F1-score grounding metric.

Language Model Size Comparison: We evaluate T5's performance across various sizes, presenting results in Table 4. T5 large outperforms T5 small, aligning with expectations due to its larger parameter count. In addition, our observations yield two primary findings that emerge through two-stage tuning: (1) Two-stage tuning enhances T5 large's performance, particularly improving token-level matching overlap such as nouns and verbs (Table 5). This improvement, consistent across T5 sizes, suggests enhanced visual content recognition, attributed to weight initialization and warming-up of the projection matrix. (2) Despite the close total performance,T5 large outperforms T5 small in causal questions by 2%-3% (Table 6), hinting at the potential for guiding language models in recognizing causal relationships.

4.3.3 Causal and Temporal Inference Abstraction

In this section, we present the outcomes of our two methods employed to abstract the causal and temporal relationships embedded within the events and entities within a video, with the ultimate aim of generating inferential questions.

Video:

Ground Truth Questions:
1: how did the girl keep her hair away from her face?
2: what did the girl do after she stood up at the beginning of the video?
3: where is this video taken?
4: what did the man gestured to the girl near the start of the video?
5: what did the girl do after the man pat the pillow?
6: why was the man looking left and right at the beginning of the video?
7: what did the girl do after she sat down?
8: why did the man touch the girl s leg when she sit beside him?
9: why did the man pat the pillow?
10: what did the man do after he adjusted the girl s leg?
11: how did the man see clearly?
12: why is the lady in green smiling?
13: why did the man lie backwards at the end of the video?

BLIP2: (***48 matching overlap***)
"1": "how did the girl kept her hair out of her face?",
"2": "what does the girl do after the man touches her at the end?",
"3": "where is this video taken?",
"4": "how does the man in white hold the child s hand?",
"5": "why did the man in white squat down in the middle of the video?",
"6": "what does the man in white do after the girl sits down?",
"7": "what does the girl do after looking at the man for a while at the end?",
"8": "why did the girl put her leg on the table in the middle of the video?"
"9": "why did the man in black stretch his hand out at the end of the video?",
"10": "what did the man do after he looked at the girl?",
"11": "how did the man in white ensured he can see the girl clearly?",
"12": "why did the man laugh at the girl?",
"13": "how did the man in black react when the girl s hands were pushed to him?"

CLIP: (***52 matching overlap***)
"1": "how did the girl kept her hair out of her face?",
"2": "what does the girl do after the man puts her back on the sofa?",
"3": "where is this video taken?",
"4": "how does the man hold the child s hand?",
"5": "why did the man in red hold the girl s hand?",
"6": "what does the man do after the girl sits on the sofa?",
"7": "what did the girl do after looking at the man?",
"8": "why did the girl bend down when she is standing?",
"9": "why did the man point to the table at the end of the video?",
"10": "what did the man do after he looked at the girl?",
"11": "how did the man see the girl clearly?",
"12": "why did the man laugh at the girl?",
"13": "why did the man pull the girl s back?"

Figure 3: Visual encoder CLIP and BLIP2 performance. Yellow scopes represent matching overlap with ground truth questions. Red scopes represent the more details recognized by the BLIP model compared with the CLIP model.

model	B	RL	M	C	Grounding
T5 Small One Stage	0.1564	0.4216	0.3594	1.0366	0.3505
T5 Small Two Stage	0.1559	0.4181	0.3594	1.002	0.3453
T5 Large One Stage	0.1459	0.4025	0.3459	0.9449	0.3249
T5 Large Two Stage	**0.1572**	**0.4281**	**0.3634**	**1.0657**	**0.3573**

Table 4: Difference Language Size Performance. T5 small has *60M* parameters, with total **135M** parameters for a whole framework, T5 large has *770M* parameters, with total **917M** parameters for a whole framework. B is BLEU, RL is ROUGEL, M is METEOR, C is CIDEr, and Grounding is the F1-score grounding metric.

model	NN	WRB	VBZ	VBD	VB	JJ	VBG	WP	PRP
T5 Small One Stage	4199	**2692**	1121	1154	713	504	248	1038	220
T5 Small Two Stage	4287	2640	1268	**1184**	643	**533**	228	**1091**	**221**
T5 Large One Stage	3927	2664	**1429**	947	719	467	227	1048	187
T5 Large Two Stage	**4478**	2655	1379	1078	**777**	517	**277**	1024	207

Table 5: Number of matching overlaps for various word types based on Spacy about the difference language model sizes. NN means noun, singular or mass, WRB means wh-adverb, VBZ means verb, 3rd person singular present, VBD means verb, past tense, VB means verb, base form, JJ means adjective, VBG means verb, gerund or present participle, WP means wh-pronoun, personal, PRP means pronoun, personal.

Vision Projection Matrix Comparison explores projection matrix techniques, revealing unexpected trends shown in Table 7. Contrary to expectations, the method directly concatenating CLIP encoder and language embeddings ("Video MLP" in Table 7) outperforms that employing the addition of MLP layers to each frame before concatenating with the language embedding ("Video 16to5 MLP" in Table 7), including grounding metrics on causal and temporal questions (Table 8). Findings underscore that the blind proliferation of MLP layers, even on individual frames, *fails to*

Model	C G-Pre	C G-Re	C G-F1	T G-Pre	T G-Re	T G-F1
T5 Small two stage	0.3096	0.3078	0.3087	0.3625	0.3357	0.3486
T5 large two stage	**0.3333**	**0.3115**	**0.3221**	**0.3767**	**0.3374**	**0.3560**

Table 6: Grounding evaluation performance of different sizes of T5 models with the two-stage tuning method in causal and temporal inference. C G represents the grounding metric of causal questions. T G represents the grounding metric of temporal questions. "Pre" represents precision. "Re" represents recall. "F1" represents the F1 score.

capture inferential relationships in visual content.

Model	B	RL	M	C	Grounding
Video MLP	**0.1564**	**0.4216**	**0.3594**	**1.0366**	**0.3505**
Video 16to5 MLP	0.1549	0.4170	0.3574	0.9722	0.3415

Table 7: Vision Projection Matrix Performance. Both experiments are conducted with a CLIP image encoder and T5-small. Video MLP means the vision embedding would be processed by a MLP layer and video 16to5 MLP means we add 16 fine-grained MLP for the frames of the video input. B is BLEU, RL is ROUGEL, M is METEOR, C is CIDEr, Grounding is the F1-score grounding metric.

Model	C G-Pre	C G-Re	C G-F1	T G-Pre	T G-Re	T G-F1
Video MLP	**0.3204**	**0.3072**	**0.3137**	**0.3695**	**0.3331**	**0.3503**
Video 16to5 MLP	0.3028	0.3014	0.3021	0.3589	0.3316	0.3447

Table 8: Vision Projection Matrix Grounding Performance in Causal and Temporal Inference. C G represents the causal grounding metric. T G represents the temporal grounding metric. "Pre" represents precision. "Re" represents recall. "F1" represents the F1 score.

Frame Comparison Based on CLIP evaluates two frame comparison methods using the CLIP-based approach. The summarized evaluations are presented in Table 9 along with an illustrative example shown in Appendix Figure A1, yielding several noteworthy findings:

1. While slightly behind direct vision embedding concatenation (Video MLP) across all evaluation metrics in Table 9, the global frame method with only 73M parameters is less than the direct concatenation approach (135M). In addition, the global frame comparison method outperforms the baseline (Random Select) on all metrics in Table 9 and has a substantial 20% boost compared to its baseline in causal and temporal questions (Table 10). Moreover, the global frame method excels in the direct concatenation approach in the grounding metrics of temporal questions within videos.

2. The local frame comparison method yields inferior results compared to its global counterpart across all evaluation metrics in Table 9. Aligning these findings with the performance of random selection, **we argue that maintaining a consistent relationship between input frames during both training and inference phases is pivotal for enabling the language model to deduce relationships between events and entities within videos effectively.** The method of random selection introduces *the highest level* of inconsistency compared to global and local frame comparison methods between training and inference due to its reliance on random frame selection throughout both phases. Additionally, an examination of CLIP frame selection based on questions and answers in the *local frame comparison method* reveals certain limitations. While instances of accurate frame selection aligned with questions and answers are observed, inherent challenges persist (Challenge examples are provided in Figure 4): (1) Descriptive questions such as "Where is this video happening?" often fail to pinpoint a specific frame, leading to varied frame selections by the CLIP model for identical questions. (2) Given that some videos within the NExT-QA dataset (Xiao et al., 2021) last 1 to 2 minutes, with only 16 available frames for video input, the CLIP model tends to select frames with similar content regardless of chronological time order if the event described in the question has not been captured by the 16 frames. These issues exacerbate inconsistencies and disorderliness in input frames between training and inference, resulting in *comparatively poorer performance of local frame comparison method* compared to the *global frame comparison method*. In conclusion, the global frame method introduces the least inconsistency, consistently measuring cosine similarity and selecting the least similar frame pair for language model input.

3. To further support our argument, we conduct an additional experiment where the initial and final (1&16) frames are consistently selected as the video input for the language model, as outlined in the fifth row of Table 9. Remarkably, the performance of this fixed selection method, while slightly distinct, consistently trails behind that of the global frame selection across all evaluation metrics except causal grounding metrics. This observation lends additional support to our argument, reinforcing the validity of our premise. Moreover, it opens a promising avenue for future exploration — **seeking methods that improve consistent relationships with frame-based techniques**.

model	B	RL	M	C	Grounding
All 16 frames (Video MLP)	**0.1564**	**0.4216**	**0.3594**	**1.0366**	**0.3505**
Two frames (Random Select)	0.0796	0.3128	0.2173	0.2520	0.2082
Two frames (Global Frame Comparison)	0.1538	0.4165	0.3578	1.007	0.3417
Two frames (Local Frame Comparison)	0.1315	0.3946	0.3316	0.8576	0.3095
Two frames (Fixed Selection) Frame 1&16	0.1526	0.4161	0.3549	0.9745	0.3407

Table 9: Frame Comparison Performance. "Video MLP" means the vision embedding would be processed by a MLP layer; "Random Select" means we randomly select two frames embedding within a video as the vision input. B is BLEU, RL is ROUGEL, M is METEOR, C is CIDEr, and Grounding is the F1-score grounding metric.

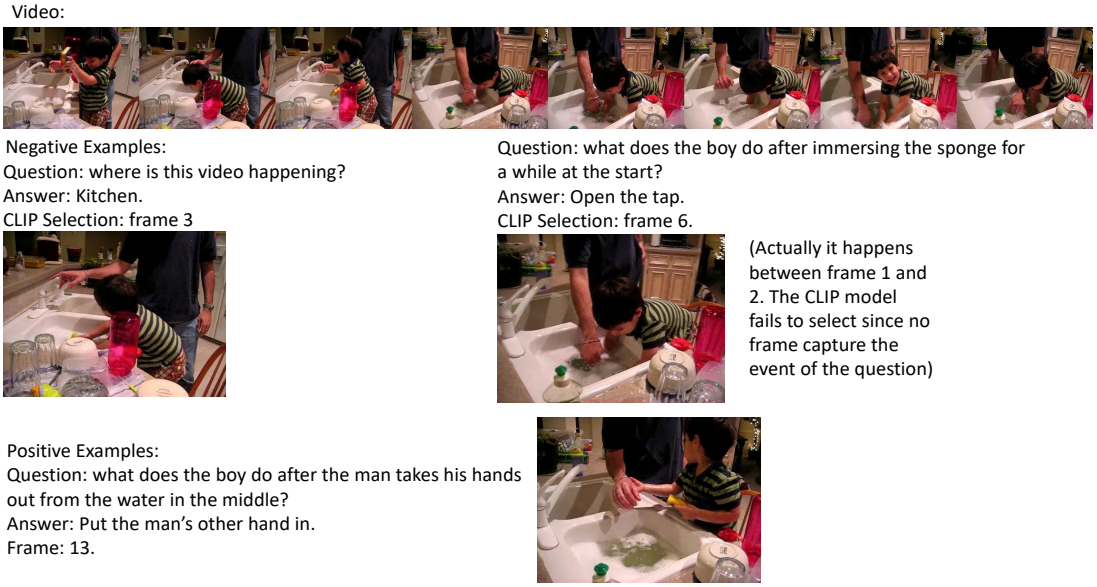


Figure 4: CLIP Selection Performance. The negative example on the left explains the inherent Challenge 1 and another negative example on the right explains the inherent Challenge 2. The positive example displays the correct frame selection.

Model	C G-Pre	C G-Re	C G-F1	T G-Pre	T G-Re	T G-F1
Video MLP	**0.3204**	0.3072	**0.3137**	0.3695	0.3331	0.3503
Random Select	0.3121	0.2340	0.2674	0.2191	0.1375	0.1689
Global Frame Comparison	0.3089	**0.3074**	0.3081	**0.3817**	**0.3509**	**0.3656**

Table 10: Global Frame Comparison Grounding Performance in Causal and Temporal Inference. C G represents the grounding metric of causal questions. T G represents the grounding metric of temporal questions. "Pre" represents precision. "Re" represents recall. "F1" represents the F1 score.

5 Conclusion

This paper bridges the gap in aligning machine-generated visual questions, focusing on inferential questions in video VQG. Our framework utilizes pre-trained models to enhance event-entity inferential relationships and question generation. We additionally introduce a grounding metric and propose techniques for causal and temporal abstraction. Through extensive experiments, we achieve significant improvement across all metrics, highlighting our framework's efficacy in promoting visual content recognition. We underscore the importance of consistent relationships between input frames during training and inference for event-entity relationship inference. This research opens a promising avenue for future work, focusing on methods to enhance consistent frame-based relationships in causal and temporal video inference.

Limitation

We employ the T5 encoder-decoder language model because of its excellent performance within the 500M to 1B parameter scope and limited GPUs. Future research could lie in exploring the inferential video VQG task with larger parameters and decoder-only language model structures. In addition, future research could separately research causal and temporal relationships between entities within videos. We attempted some methods that had negative effects on our framework and experiments. These include applying contrastive learning and visual-semantic arithmetic inferential relations. Details and results of these methods are provided in the Appendix, offering references for future research.

Acknowledgments

This work originated from a dissertation at the University of Edinburgh. We must thank Yijun Yang, Hanxu Hu, Danyang Liu, and Pizhen Chen for giving valuable suggestions on this work. We are grateful to the anonymous reviewers for their advice to further clarify the novelty of this work.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Shih-Han Chan, Tsai-Lun Yang, Yun-Wei Chu, Chi-Yang Hsu, Ting-Hao Huang, Yu-Shian Chiu, and Lun-Wei Ku. 2022. Let's talk! striking up conversations via conversational visual question generation. *arXiv preprint arXiv:2205.09327*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. 2021. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

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

Pin Jiang and Yahong Han. 2020. Reasoning with heterogeneous graph alignment for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11109–11116.

Zaid Khan, Vijay Kumar BG, Samuel Schulter, Xiang Yu, Yun Fu, and Manmohan Chandraker. 2023. Q: How to specialize large vision-language models to data-scarce vqa tasks? a: Self-train on unlabeled images! In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15005–15015.

Khushboo Khurana and Umesh Deshpande. 2021. Video question-answering techniques, benchmark datasets and evaluation metrics leveraging video captioning: A comprehensive survey. *IEEE Access*, 9:43799–43823.

Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2019. Information maximizing visual question generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2008–2018.

Che-Hao Lee, Tzu-Yu Chen, Liang-Pu Chen, Ping-Che Yang, and Richard Tzong-Han Tsai. 2018. Automatic question generation from children's stories for companion chatbot. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 491–494. IEEE.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xiao Lin and Devi Parikh. 2016. Leveraging visual question answering for image-caption ranking. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 261–277. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. *arXiv preprint arXiv:1603.06059*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

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

Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2022. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17928.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Nihir Vedd, Zixu Wang, Marek Rei, Yishu Miao, and Lucia Specia. 2021. Guiding visual question generation. *arXiv preprint arXiv:2110.08226*.

Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. 2021. Rethinking infonce: How many negative samples do you need? *arXiv preprint arXiv:2105.13003*.

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

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Min-Hsuan Yeh, Vincent Chen, Ting-Hao Huang, and Lun-Wei Ku. 2022. Multi-VQG: Generating engaging questions for multiple images. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 277–290, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhenjie Zhao, Yufang Hou, Dakuo Wang, Mo Yu, Chengzhong Liu, and Xiaojuan Ma. 2022. Educational question generation of children storybooks via question type distribution learning and event-centric summarization. *arXiv preprint arXiv:2203.14187*.

A Appendix

A.1 Causal and Temporal Inference Abstraction

Videos:

Ground Truth Questions:
1: why is the man in green holding onto the shoe as the man in white is cutting it?
2: what does the man in white do after holding the shoe stably?
3: why did the man in green point his hand at the man in white while he is talking?
4: why did the man in white pick up a knife after changing place with the man in green?
5: why does the man in green hold a shoes in his hand at the start?
6: what does the man in white do after cutting the shoes for a while?
7: why did the man in green hold up the shoe as he is speaking?
8: how do the men appear while cutting the shoes?
9: why are the men looking down at the shoe while the man in white is cutting it?
10: why did the man in white stand beside the man in green while he is speaking?
11: why did the man in white move behind after the man in green put down the shoe on the table?

Global Frame Comparison: (***42 matching overlap***)
"1": "why did the man in black hold onto the man in black s hand when he walks ?",
"2": "how did the man in black show that he wants to cut the glass?",
"3": "why did the man in black look at the man in black when he walks to the man in black ?",
"4": "why did the man in black bend down at the start ?",
"5": "why did the man in black put his hand on the glass in the middle of the video?",
"6": "why did the man in black point to the man in black when he is speaking?",
"7": "what did the man in black do after he walked to the man in black ?",
"8": "how does the man in black look while talking ?",
"9": "why did the man in black hold onto the bottle when he walks?",
"10": "why did the man in black walk back to the man in black after he finished talking?",
"11": "why did the man in black change his position after he sat down"

Video MLP: (***39 matching overlap***)
"1": "why did the man in black hold onto the man in red s hand when he walks?",
"2": "what did the man in black do after he pointed at the grass at the start?",
"3": "why did the man in black look at the man in black after he stops walking?",
"4": "why did the man in black bend down at the start?",
"5": "why did the man in black hold the stick in his hand?",
"6": "why did the man in black point to the man in black when he is talking?",
"7": "why did the man in black bend down at the start?",
"8": "how does the man in black look while talking?",
"9": "why is the man in black holding the stick?",
"10": "why did the man in black walk towards the man in green after he finished talking?",
"11": "why did the man in black move backwards after he has finished talking?"

Figure A1: Frame Comparison Performance. Yellow scopes represent matching overlap with ground truth questions. Red scopes represent more details recognized by the frame comparison method compared with the Video MLP method.

A.2 Negative Methods for Causal and Temporal Inference Abstraction

A.2.1 Contrastive Learning Based on Frame Comparison

Contrastive Learning on Unifying Vision and Language Embedding aims to leverage the nuanced interplay between video frames using contrastive learning. The infoNCE loss function (Oord et al., 2018) is employed for contrastive learning (Wu et al., 2021), maximizing the lower bound of mutual information between pairs of variables. The core framework encompasses a relevance function such as cosine similarity, represented as $f(\cdot, \cdot)$, where each positive sample (x^+, c) is linked with a set of k randomly chosen negative samples denoted as $(x_1^-, c), (x_2^-, c), ..., (x_k^-, c)$. Then, the InfoNCE loss function $\mathcal{L}_k$ is formulated as follows:

$$\mathcal{L}_k = -\log\left(\frac{e^{f(x^+,c)}}{e^{f(x^+,c)} + \sum_{i=1}^{k} e^{f(x_i^-,c)}}\right) \quad (5)$$

Positive samples are derived from two frame pairs: the global contradictory frame pair and the local contradictory frame pair, similar to the methods in the Contradictory Frame Comparison Section. The remaining frames, paired with the second frame from each contradictory set, serve as negative samples. These positive and negative samples, along with the second frame's embedding, are used in the infoNCE loss formula. The contrastive learning loss is integrated with the pre-trained language model loss, defining the total loss function. Formally, the total loss function was defined as:

$$\mathcal{L}_{Total} = \mathcal{L}_{language\ model} + \mathcal{L}_k \quad (6)$$

A.2.2 Visual-Semantic Arithmetic Inferential Relation

Visual-Semantic Arithmetic Inferential Relation Abstraction aims to capture relationships between frames within a video by subtracting frame embeddings. Drawing inspiration from recent findings (Tewel et al., 2022; Goh et al., 2021) on the CLIP multi-modal representation, we develop a loss function which is adapted to guide the language model in recognizing relationships, especially causal and temporal ones. Specifically, we first compute the relevance of frames for potential tokens at length i. Top K token candidates are selected, while the remaining tokens are assigned zero potential to enhance computational efficiency. These candidate sentences, denoted as $s_i^k = (x_1, ..., x_{i-1}, x_i^k)$, correspond to the k-th candidate token and are matched against the frame I. It is pertinent to highlight that the context tokens $x_1, ..., x_{i-1}$ are constant for the current token x_i^k. Subsequently, the frame potential of the k-th token is computed as:

$$D_i^k \propto \exp\left(\frac{F_{cos}(E_{Text}(s_i^k), E_{frame}(I))}{\tau_c}\right), \quad (7)$$

Here, F_{cos} represents the cosine distance between CLIP's embeddings of the text (E_{Text}) and the frame (E_{Image}). The hyperparameter $\tau_c > 0$ is a temperature parameter that adjusts the sharpness of the target distribution. In our experiments, it was set to 0.05. Notably, the frame embedding E_{Image} emerges from subtracting the CLIP image embeddings of two frames. Subsequently, the CLIP loss materializes as the cross-entropy loss between the frame potential distribution and the target distribution of the next token x_{i+1} derived from the language model:

$$\mathcal{L}_{CLIP} = CE(D_i, x_{i+1}). \quad (8)$$

This loss encourages the language model to discern relationships between frames, fostering causal and temporal inferences. The total loss function combines the language model loss and the CLIP loss:

$$\mathcal{L}_{Total} = \mathcal{L}_{language\ model} + \mathcal{L}_{CLIP} \quad (9)$$

A.3 Experiment Results on Negative Methods for Causal and Temporal Inference Abstraction

model	B	RL	M	C	Grounding
Global Frame Comparison baseline	0.1538	**0.4165**	0.3578	1.007	0.3417
Global Frame Comparison Contrast	**0.1555**	0.4164	**0.3601**	**1.010**	0.3383
Local Frame Comparison Contrast	0.1531	0.4165	0.3555	1.001	**0.3426**

Table A1: Contrasting Learning Performance. The baseline is the "Global Frame Comparison" shown in Table 5. B is BLEU, RL is ROUGEL, M is METEOR, C is CIDEr, and Grounding is the F1-score grounding metric.

A.3.1 Experiment Results on Contrastive Learning Based on Frame Comparison

Contrastive Learning Based on Frame Comparison evaluates two contrasting learning methods rooted

in global frame comparisons, summarized in Appendix Table A1. Surprisingly, both global frame contrast and local frame contrast methods outperform the baseline in specific metrics, showcasing the potential of contrastive learning in enhancing the language model's ability to discern nuanced details within videos, such as characters, colours, verbs, and tense, as illustrated within Appendix Figure A2's red scope. Despite the marginal overall performance difference, contrasting learning proves beneficial for the language model in understanding video content and generating inferential questions, particularly concerning temporal relationships, shown in Appendix Table A2 and Appendix Table A3. However, the similarity in performance raises considerations about the limited negative sample pool and the constrained parameters of the T5 small model, affecting the model's ability to differentiate between positive and negative samples during contrastive learning. This observation highlights the need for a more extensive negative sample pool and suggests potential limitations in the model's capacity to encompass comprehensive knowledge for effective contrastive learning in continuous video data.

A.3.2 Experient results on Visual-Semantic Arithmetic Inferential Relation

Visual-Semantic Arithmetic Inferential Relation reveals that the visual-semantic arithmetic method's performance closely resembles the baseline approach of directly concatenating vision embeddings, detailed in Table A4. This suggests that supplementing the visual-semantic arithmetic with CLIP loss may not significantly enhance performance. A comparison of questions generated by two frame selection techniques indicates similarities and disparities, with examples presented in Appendix Figure A3. Examination of generated questions in causal and temporal types, along with matching overlap levels with the baseline, is detailed in Appendix Table A5. However, the visual-semantic arithmetic method outperforms in temporal questions, exhibiting a 1-2% increase compared to direct vision concatenation, particularly excelling in recognizing time adverbs. Despite its effectiveness, the method's reliance on multi-model concatenation may fall short in enabling the language model to comprehensively discern the complete spectrum of visual relationships within contrasting frame pairs in a video, as suggested by examples in Appendix Figure A4.

model	NN	WRB	VBD	VBZ	VB	JJ	VBG	WP	PRP
Global Frame Comparison baseline	4166	2571	981	**1489**	776	503	**345**	1131	**247**
Global Frame Comparison Contrast	**4222**	2553	**1157**	1332	592	**558**	224	**1155**	233
Local Frame Comparison Contrast	4196	**2588**	1122	1310	**823**	530	225	1136	244

Table A2: Number of matching overlap for various word types based on Spacy about the frame contrasting methods. NN means noun, singular or mass, WRB means wh-adverb, VBZ means verb, 3rd person singular present, VBD means verb, past tense, VB means verb, base form, JJ means adjective, VBG means verb, gerund or present participle, WP means wh-pronoun, personal, PRP means pronoun, personal.

model	C G precision	C G recall	C G F1-score	T G precision	T G recall	T G F1-score
Global Frame Comparison	0.3089	**0.3074**	**0.3081**	0.3817	0.3509	0.3656
Global Frame Comparison Contrast	**0.3138**	0.2960	0.3046	0.3562	0.3383	0.3470
Local Frame Comparison Contrast	0.3010	0.2939	0.2974	**0.3972**	**0.3599**	**0.3776**

Table A3: Contrasting Learning Methods Evaluation Performance in Causal and Temporal Inference. C G represents the causal grounding metric. T G represents the Temporal causal grounding metric.

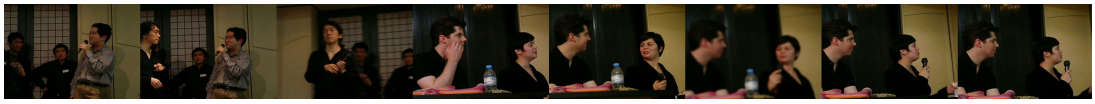
model	B	RL	M	C	Grounding
Video MLP	0.1564	**0.4216**	0.3594	**1.0366**	**0.3505**
CLIPloss top word 100	**0.1568**	0.4184	**0.3602**	1.0359	0.3460

Table A4: Visual-semantic arithmetic inferential performance. Video MLP represents the direct vision concatenation method. CLIPloss represents the visual-semantic arithmetic method. B is BLEU, RL is ROUGEL, M is METEOR, C is CIDEr, and Grounding is the grounding metric.

model	C G precision	C G recall	C G F1-score	T G precision	T G recall	T G F1-score
Video MLP	**0.3204**	**0.3072**	**0.3137**	0.3695	0.3331	0.3503
CLIPloss top word 100	0.3107	0.3061	0.3084	**0.3828**	**0.3433**	**0.3620**

Table A5: Visual-semantic Arithmetic Evaluation Performance in Causal and Temporal Inference. C G represents the causal grounding metric. T G represents the Temporal causal grounding metric.

Videos:



Ground Truth Questions:
1: what did the lady in black do after the man next to her gave her a microphone?
2: how did the lady in black reacted when the man in black beside her passed her the microphone?
3: what is the man with white tag on shirt do while man in stripes speaking?
4: why did the man in black with tied up hair turned backwards after he received the microphone?
5: what is the lady in black doing with her hands as she spoke into the microphone at the end of the video?
6: what did the man in grey do after he finished his speech?
7: what did the man in black in front of the man in grey do before the man in grey passed him the microphone?
8: why did the man in black with tied up hair walked towards the man in grey in the middle of the video?
9: why is the lady in black moving her hands at the end of video?
10: why did the lady in black face the man in black beside her before she started talking into the microphone?

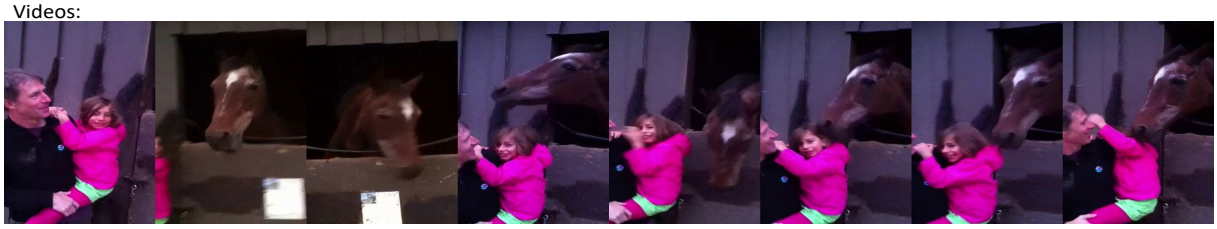
Global Frame Contrast Learning: (***58 matching overlap***)
"1": "what does the lady in black do after the man in black points at her at the start?",
"2": "how did the man in black react when the man in black was talking?",
"3": "what did the man in black do as the man in white was talking?",
"4": "why did the man in black walk away after he finished talking?",
"5": "why did the man in black move his hands away from the lady in white?",
"6": "what did the man in black do after he finished speaking?",
"7": "what did the man in black do after the man in grey walked away?",
"8": "why did the man in black walk towards the man in black?",
"9": "why did the man in black raise his hands in the air at the end of the video?",
"10": "what did the lady in black do after she turned to face the man in black?"

Global Frame Baseline: (***40 matching overlap***)
"1": "what does the man in blue do after the man in blue points at him at the start?",
"2": "what does the man in blue do after he finishes talking?",
"3": "what did the man in blue do after he walked away from the man in blue?",
"4": "why did the man in blue walk away after he walked away?",
"5": "why did the man in blue move his hand towards the lady in blue at the end of the video?",
"6": "what did the man in black do after he finished talking?",
"7": "what did the man in black do after the man in grey walked away at the end of the video?",
"8": "why did the man in blue walk towards the man in blue?",
"9": "why did the man in black move his hands as he speaks?",
"10": "what did the man in blue do after he pointed at the man in blue?"

Local Frame Contrast Learning: (***47 matching overlap***)
"1": "what does the man in black do after the man in black starts speaking?",
"2": "what did the man in black do after he took the photo?",
"3": "what does the man in black do as the man in black was talking?",
"4": "why did the man in black walk away after he talked to the man in black?",
"5": "why did the man in black move his hand towards the lady in black?",
"6": "what did the man in black do after he finished singing?",
"7": "what did the man in black do after the man in grey walked away?",
"8": "why did the man in black walk towards the man in black?",
"9": "why did the man in black move his hands as he speaks?",
"10": "what did the man in black do after he walked to the man in black?"

Figure A2: Contrast Learning Performance. Yellow scopes represent matching overlap with ground truth questions. Red scopes represent more details recognized by the frame contrasting methods compared to the global frame comparison method.

Videos:



Ground Truth Questions:
1: where is the man and the girl?
2: how does the man react when the horse plays with the girl?
3: what does the girl do after patting the horse?
4: how does the girl in pink react when the horse licks her?
5: what does the horse do after the girl pats it in the middle of the video?
6: how does the girl react when the horse turns towards her the first time?
7: does the girl seem more scared or excited to play with the horse?
8: what does the horse do after it turns back to the girl the second time?
9: why does the girl move her head away from the horse at the end of the video?
10; what does the man do when the horse plays with the girl?
11: what is the animal show in the video?
12: what does the girl do after tucking her hair behind her ear?

Visual-semantic Arithmetic Method: (***51 matching overlap***)
"1": "where is this video taken?",
“2": "how does the girl react when the man is playing with her?",
“3": "what did the girl do after the man walked away?",
“4": "how does the girl react when the man is playing with her?",
“5": "what does the girl do after the horse approaches her at the end?",
“6": "how does the girl react when the man is playing with her?",
“7": "why did the girl start jumping when the horse is near her?",
“8": "why did the girl put her hand on her face when the horse approached her?",
“9": "why did the girl bend down at the end of the video?",
“10": "how does the man support the girl as she stands on the horse?",
“11": "what is the animal shown in the video?",
“12": "what does the girl do after the man puts her down?"

VideoMLP Baseline: (***44 matching overlap***)
“1": "where is this place?",
“2": "how does the girl react after the horse jumps up?",
“3": "what does the girl do after the man approaches her at the end?",
“4": "how does the girl react after the horse jumps up?",
“5": "how does the dog show affection towards the girl?",
“6": "how does the girl react after the horse jumps up?",
“7": "why did the girl start jumping when the horse approached her?",
“8": "why did the girl put her hand on the horse after the horse jumps up?",
“9": "why did the girl run towards the horse after the horse jumped up?",
“10": "how does the man ensure the girl does not fall?",
“11": "what animal is shown in the video?",
“12": "what does the girl do after the man starts to approach her at the start?"

Figure A3: Visual-semantic arithmetic method performance. Yellow scopes represent matching overlaps with ground truth questions. Red scopes represent more details recognized by the visual-semantic arithmetic method.

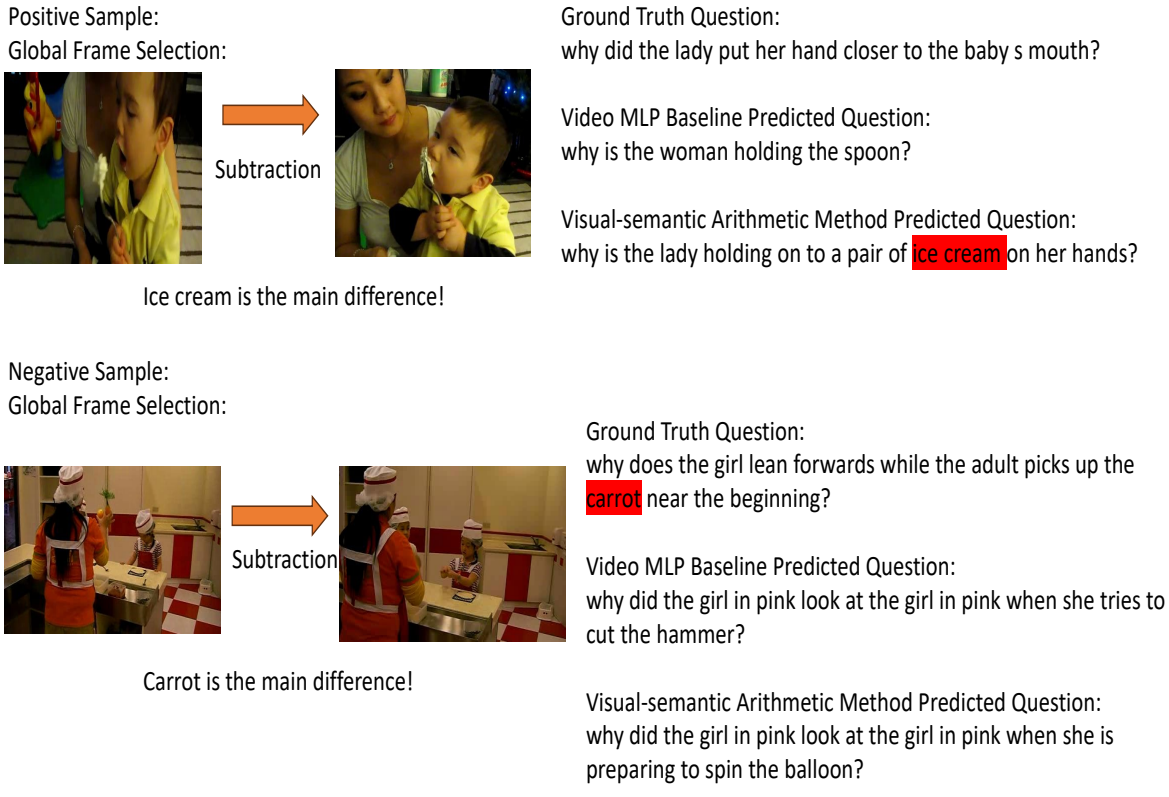


Figure A4: The effectiveness of the Visual-semantic arithmetic method: check if the language model could recognize the difference between two frames.