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

Zhanghao HU

Master of Science

Artificial Intelligence

School of Informatics

University of Edinburgh

2023

Abstract

Visual Question Generation (VQG) is a field at the crossroads of visual and language learning, impacting broad domains like education, medicine, social media, and e-commerce. Existing pre-trained models have excelled by fusing vision and language embeddings. Yet, they predominantly focus on fact-based queries using image pairs, disregarding human-like thinking that encompasses causal and temporal connections in videos. Moreover, most pre-training methods demand substantial computational resources. Leveraging relations between various pre-trained models in multi-modal learning particularly in the domains of video remains an under-explored avenue.

This study addresses the research gap in generating inferential questions concerning causal and temporal inference for video VQG. We introduce a novel framework that employs vision-text matching pre-trained models to guide language models in recognizing event-entity relationships within videos. This facilitates the generation of pertinent inferential questions involving causal and temporal inferences. Our approach's efficacy is demonstrated on NExT-QA, a dataset for causal and temporal inference in visual question answering. Experimental results confirm the success of our method in leading the pre-trained language model recognize the video content. We also present a series of methodologies for abstracting causal and temporal relationships between events and entities. Comprehensive analysis unveils the potential of our methods and point out the directions for future exploration.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Zhanghao HU*)

Acknowledgements

Time is a river with many eddies. Edinburgh is the most important stop of my study abroad journey entirely throughout my bachelor's and master's life. The months have been challenging, I have only been able to finish this project because of the dedicated support I have received from so many people and friends along the way.

First and foremost, I would like to express my strongest gratitude to my supervisor Frank Keller for his continuous and invaluable guidance throughout the project. He is not only a conscientious supervisor, but a knowledgeable teacher, and an earnest friend to me. I am definitely fortunate to discuss the project proposal with him and have him as my supervisor.

I would like to thank my friends, Mr Yijun Yang, Mr Junjie Xu, Mr Hanxu HU and Mr Ningyuan Shan, for their academic help and importantly listening to me and relieving stress.

Then I would like to thank my parents for their encouragement, patience, and love throughout my life journey. My mom is always the best listener whenever I am down or lost. I never felt alone when she was around. Though my dad is not good at expressing himself, I can still feel he is loving me in his own way.

My final thank goes to my girlfriend, Miss Wenfei Ding. Throughout my master's life, I was usually under worry and pressure, but she often listened to my countless complaints and always comforted me.

Table of Contents

Chapter 1

Introduction

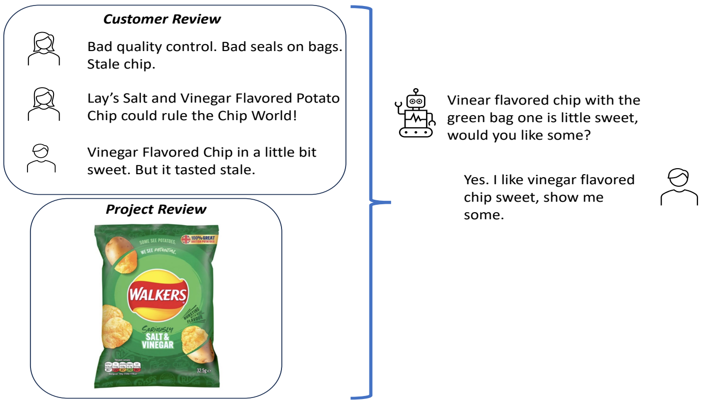
1.1 Motivation

Visual Question Generation (VQG) has emerged as a significant research area in multi-modal learning between vision and language since its inception in 2016 [34]. Its impact spans various domains like education [69], social media [66], and human-computer interaction [25]. An Example with VQG in e-commerce is shown in Figure 1.1. Currently, most questions in traditional question-answering datasets yield factoid answers[1] [66]. These answers do not align with human thinking, as they are directly derived from visual content. For instance, asking ”Was anyone injured in the crash?” after viewing an image of a car accident is uninteresting and obvious. On the contrary, inferential questions, particularly those related to causal inference and temporal inference, like ”Why do these drivers have accidents in the middle of intersections?” or ”What will the police do after the crash?”, better reflect human thoughts as they provide valuable insights that cannot be directly answered using visual content.

Despite the progress in VQG, no research has yet explored inference and reasoning that align with human thinking. Moreover, unlike singular images, videos possess the capability to depict relationships between events and entities. Therefore, our work focuses on two classical reasoning: causal inference and temporal inference, aiming to bridge this gap and introduce a new challenge in the field of video visual question generation. Examples of causal and temporal questions are provided in Fig 1.2.

Meanwhile, the rapid rise and impressive capabilities of visual transformers [9] have led to their widespread use in various multi-modal learning tasks that bridge the

[1] Factoid question answering: Questions directly inquire about visual facts based on the provided visual information.



Bad quality control. Bad seals on bags. Stale chip.

Lay’s Salt and Vinegar Flavored Potato Chip could rule the Chip World!

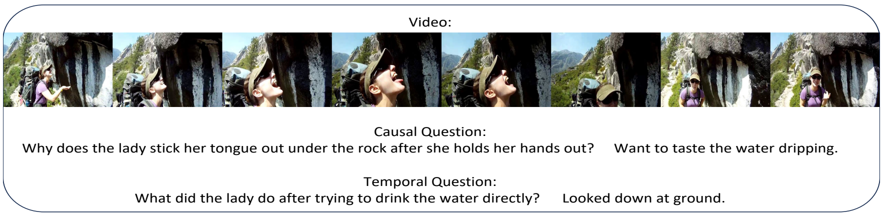
Vinegar Flavored Chip in a little bit sweet. But it tasted stale.

Vinear flavored chip with the green bag one is little sweet, would you like some?

Yes. I like vinegar flavored chip sweet, show me some.

Figure 1.1: Example of visual question generation in E-commerce. The system could guide and recommend customers to their favourite products according to the image and customer reviews.

gap between vision and text. Notably, methods of pre-training large vision transformer models have been successfully applied in tasks such as image captioning [30], visual question answering [18][43], and visual grounding [41]. However, it is essential to acknowledge that these pre-training methods come with certain drawbacks. They are computationally intensive and demand high-quality GPUs for training. For instance, one of the pre-trained models [43] required a staggering 592 V100 GPUs and took 18 days to complete the training process.



Why does the lady stick her tongue out under the rock after she holds her hands out? Want to taste the water dripping.

What did the lady do after trying to drink the water directly? Looked down at ground.

Figure 1.2: Examples for causal and temporal questions.

The potential of leveraging and guiding existing pre-trained models is often overlooked, especially in generative tasks like question generation. Building upon the inspiration from prior work [37], we aim to delve into the realm of generating inferential questions by harnessing the power of pre-trained vision-to-text matching models, rather than pre-training a vision model from scratch. By capitalizing on the knowledge already captured in the vision-to-text models, we can potentially expedite the question generation process and enhance the quality of generated questions.

In this study, we will investigate the effectiveness of utilizing existing pre-trained vision-text matching models and language models for the task of generating questions. Specifically, we will focus on temporal and causal inference questions, which require a deeper understanding of visual context and reasoning capabilities. By exploring this avenue, we aim to advance the field of visual question generation and shed light on the practicality of employing pre-trained models in generative tasks, opening new opportunities for more efficient and intelligent question-generation methods.

1.2 Problem Statement

Visual Question Generation (VQG) has emerged as a significant research area since its inception in 2016 [38]. Its broad impact spans diverse domains, benefiting applications in children's education [69], radiology medicine [49], social media [66], and human-machine interaction [25]. Moreover, VQG plays a critical role in advancing comprehensive multi-modal tasks that involve both vision and text, such as visual question answering [32], visual storytelling [51], and visual dataset creation [15]. This versatile technique has already been integrated into human daily life, facilitating office language ability examinations such as Duolingo [23] and language practice applications that utilize VQG to ask questions based on given images.

Despite these advancements, the current state of VQG tasks primarily revolves around single image, lacking the capacity to infer causal and temporal relationships in dynamic visual contexts. This limitation poses a critical challenge, as understanding causality and temporal dynamics is essential for deeper comprehension and inference-based questioning. The absence of temporal and causal inferential questions in video VQG restricts the ability to reason about cause-effect relationships and temporal sequences within videos.

To unlock the potential of video-based VQG tasks and advance downstream applications, it is imperative to explore the incorporation of temporal and causal attributes. By enabling machines to generate inferential questions that encompass causality and temporal dynamics within videos, multi-modal fields can be revolutionized. Video VQG tasks equipped with the exciting challenge of understanding dynamic visual contexts will surpass the limitations imposed by static image-based question generation.

1.3 Aim and Objective

The objective of our project is to bridge the gap in visual question generation by focusing on the critical aspects of temporal inference and causal inference. We frame this challenge as an inferential question generation problem. Specifically, given a set of visual information **V**, which can include multiple images or a video, and auxiliary text information **T**, representing possible answers, our aim is to explore how pre-trained language models can be leveraged to generate meaningful questions **Q** that go beyond simple factual inquiries.

To achieve this aim, our project has several specific objectives:

1. ***Generate inferential questions that surpass conventional factoid queries.*** Existing question-answering datasets predominantly consist of factoid questions that directly inquire about visual facts. However, these queries often lack depth and fail to capture the essence of causal and temporal relationships. Our objective is to produce inferential questions that better align with human thinking and understanding, as humans naturally contemplate the reasons and inferences behind visual events.

2. ***Explore the potential of utilizing pre-trained vision-text matching models and language models in recognizing visual information, particularly in videos, for generating questions focused on temporal and causal inference.*** Traditional methods such as pre-training in the visual-to-text domain often rely on time-consuming and expensive techniques, such as masked language models. By harnessing the extensive textual knowledge already present in pre-trained language models and vision-text matching models, we can guide them to generate textual output based on additional visual information.

1.4 Contribution

The contribution of visual question generation related to temporal and causal inference questions in this MSc dissertation can be outlined as follows:

1. **Proposing a Novel Framework for Video Question Generation**: The dissertation introduces a pioneering framework that leverages vision embeddings from pre-trained vision-text matching models. It guides pre-trained language models to generate inferential questions related to video content.

2. **Innovative Visual Encoder Comparison and Training Methods**: The work includes a comparison of various visual encoders, ranging from classical to state-of-the-art. Moreover, a new training method is proposed specifically designed for large pre-trained language models. The improvement demonstrates the effectiveness of the proposed framework in video question generation tasks.

3. **Introduction of a Novel Grounding Metric**: Acknowledging the limitations of general evaluation metrics, the dissertation presents a novel grounding metric. This metric aids in evaluating the extent to which the predicted questions are aligned with the content of the video. This contribution improves the reliability of the assessment process.

4. **Exploration of Causal and Temporal Inference in Videos**: The work delves into several methods to capture causal and temporal inference within a video context. The analysis offers valuable insights into how the proposed framework can effectively address questions related to causality and temporal relationships and provides a foundation for future research in the video VQG domain.

1.5 Thesis Outline

This section outlines the remainder of the dissertation:

1. **Chapter 2** unfolds the related work and background knowledge about multi-modal learning. It introduces the realm of visual question generation, pre-trained models adept at matching vision to text and pure language pre-trained models.

2. **Chapter 3** defines the task and relevant datasets of visual question generation. Moreover, it presents a novel grounding metric specifically designed to assess the relevance of generated questions with the content of the video.

3. **Chapter 4** unveils the core methodology used in the study. The chapter details the curation of model components. and further explores multi-modal fusion techniques.

4. **Chapter 5** highlights the execution of the proposed framework and presents an analysis of the obtained results, offering an understanding of their impact on visual question generation.

5. **Chapter 6** concludes the outcomes of the study and discuss potential avenues for future research. The dissertation contributes to the advancement of knowledge in this domain and inspires further exploration and innovation in the future.

Chapter 2

Background and Related Work

This chapter provides the necessary background knowledge and reviews recent relevant works that underpin this thesis. Section 2.1 presents an overview of existing research in visual question generation and explores the concepts of temporal inference and causal inference, Section 2.2 delves into recent advancements in pre-trained vision-to-text matching models and language models, elucidating how these models can be effectively utilized to replace conventional pre-training methods in multi-modal generative tasks.

2.1 Visual Question Generation

Visual question generation (VQG) is a field at the intersection of computer vision and natural language processing, where machines are trained to generate meaningful questions based on visual content. While VQG has witnessed significant progress, 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 and understanding of visual content, going beyond mere factual queries.

Since the algorithms employed in VQG tasks are diverse and dispersed, to provide a clear overview, we will summarize the tasks and their respective characteristics in Table 2.1. We will divide this section into two main parts: the first part will focus on existing VQG tasks, exploring various approaches and methodologies that have been proposed in the literature. The second part will delve into causal and temporal inference, specifically examining a related task known as video-based visual question answering (VQA), which shares similarities with our objective of generating inferential questions in the visual domain.

Task	Example	Main Models	Dataset	Inference
Signle Image VQG	[56]	Bert [8]&Transformer [54]	VQA v2.0[1]	Not
	[22]	CNN & LSTM [13]	VQA [1]	Not
Multiple Image VQG	[4]	T5 [45]	SQuAD [46]	Not
	[66]	VL-Bart & VL-T5 [6]	VIST [14] & MVQG [66]	Not
Video VQG	[52]	Faster-RCNN [47] & LSTM[13]	Anet-QA [67]& TVQA [26]	Not
	[11]	Attention & Transformer [54]	YouTube-Clips [5]	Not
Open-end Video VQA	[60]	GRU [7]	NExT-QA [60]	Causal&Temporal
Multiple-choice Video VQA	[60]	GRU [7]	NExT-QA [60]	Causal&Temporal
	[61]	QGA [61]&GCN [21]	NExT-QA [60] & MSVDQA[64]	Causal&Temporal
	[62]	Transformer [54]&GCN [21]	NExT-QA [60] &Causal-VidQA [28]	Causal&Temporal

Table 2.1: Summary of current methodologies of different tasks. GRU: Gated Recurrent Unit. QGA: Query-conditioned Graph Attention unit. GCN: Graph Convolution Network.

2.1.1 Exitsing Visual Question Generation Tasks and Techniques

The task of Visual Question Generation (VQG) was first introduced in 2016[38], aiming to generate questions based on individual images. Since then, extensive research has been conducted in this area, exploring techniques for both multiple-image VQG and video VQG. Compared to single-image VQG, multiple-image VQG and video VQG offer exciting prospects due to their potential to infer causality and temporal relationships between events and entities, making them worthy areas of investigation. As a result, this section will focus on the more promising research areas of multiple-image VQG and video VQG.

Single and Multiple Image VQG. Single image VQG involves generating questions about a single image, with recent studies exploring methods to produce specific types of questions, such as grounding or implicit questions [56] and spatial or temporal questions [22]. In contrast, multiple-image VQG is introduced by Chan et al. [4], and subsequently, Yeh et al. further advanced this research by proposing a multiple-image question generation dataset that includes summary information as the auxiliary text for each image series in 2022 [66].

Video VQG. Video VQG aims to inquire user queries in natural language based on videos, remaining relatively unexplored compared to image VQG. This is partly due to the absence of specific VQG datasets, since existing visual question-answering (VQA) datasets often provide short answers rather than complete sentences or paragraphs, which limits their utility for training video VQG tasks [16][19]. Moreover, the spatio-temporal nature of videos introduces additional complexity, making a straightforward extension of image VQG techniques insufficient for optimal results in video VQG. Currently, the models and application scenarios for video VQG lack standard categorization, but two paradigms that have shown promise in generating meaningful questions from video inputs are the encoder-decoder approach and attention networks [52][11].

2.1.2 Deriving Video Question Generation from Inferential Video Question Answering

Research on video VQG is still in its nascent stage and scattered, but there has been significant progress in video VQA tasks, particularly focusing on multiple-choice and open-end answers. However, open-end video VQA, which involves generating diverse and complex answers, remains unexplored due to its challenging nature, while multi-choice video VQA is often used to study inference-based QA beyond simple factoid questions [60][58][28]. In this section, we introduce the multi-choice video VQA and open-end video VQA and propose the transfer of inference algorithms.

Open-end Video VQA. Open-end video VQA can be categorized as classification, generation, or regression, depending on the specific datasets. It is commonly defined as a multi-class classification problem, where models classify video-question pairs into a predefined global answer set [70]. Notably, the NExT-QA dataset focuses on open-end video VQA tasks that involve causal and temporal inference [60].

Multiple-choice Video VQA. Multiple-choice video VQA presents several candidate answers for each question and requires selecting the correct one. Datasets focused on inference fall into two types: normal video QA [60] and multi-model video QA [68], where the latter of which involves resources beyond visual content. Multiple-choice VQA tasks often follow an encoder-only paradigm, where the answer decoder acts as a 1-way classifier to choose the correct answer. To achieve a comprehensive understanding of videos, graph architecture networks [21] have shown promise in inference video VQA due to their ability for effective information interaction [70] [61] [62]. The most challenging aspect lies in designing sophisticated graph structures for video representation.

In conclusion, while single-image VQG has received considerable attention, multiple-image VQG and video VQG remain less explored, with limitations in short answer generation and lack of standardized categories. Incorporating attributes like causal inference and temporal reasoning into question generation remains an unexplored area. Existing multiple-image and video VQG techniques, along with inferential video VQA, offer valuable insights and network motivation for inferential VQG tasks.

2.2 Multi-modal Generative Task with Pre-trained Models

2.2.1 Replace Learning from Scratch

With the rise in popularity and efficacy of pre-trained methods, researchers have increasingly adopted large pre-trained models to bridge the gap between vision embeddings and language embeddings in visual generative tasks, such as image captioning [30], visual question answering [18], and visual grounding [41]. These pre-trained models have shown remarkable performance by combining vision and language information to generate accurate and meaningful outputs. However, the adoption of pre-trained methods comes with the downside of high computational costs and time-consuming training, often requiring hundreds of GPUs and weeks of training time [9], making it unfeasible for many researchers with limited resources.

An alternative approach, utilizing vision-text matching pre-trained models [43][30] [29], offers a promising solution. These models have demonstrated their capability to bridge the gap between vision and language domains effectively. By leveraging the knowledge learned in these vision-text matching models, it becomes possible to guide language models in generating text outputs, thus significantly reducing the time and computation required compared to traditional pre-trained methods. To the best of our knowledge, no prior research has explored the potential of leveraging pre-trained vision and language models to guide the generation of vision-based questions, specifically those involving causal and temporal inference.

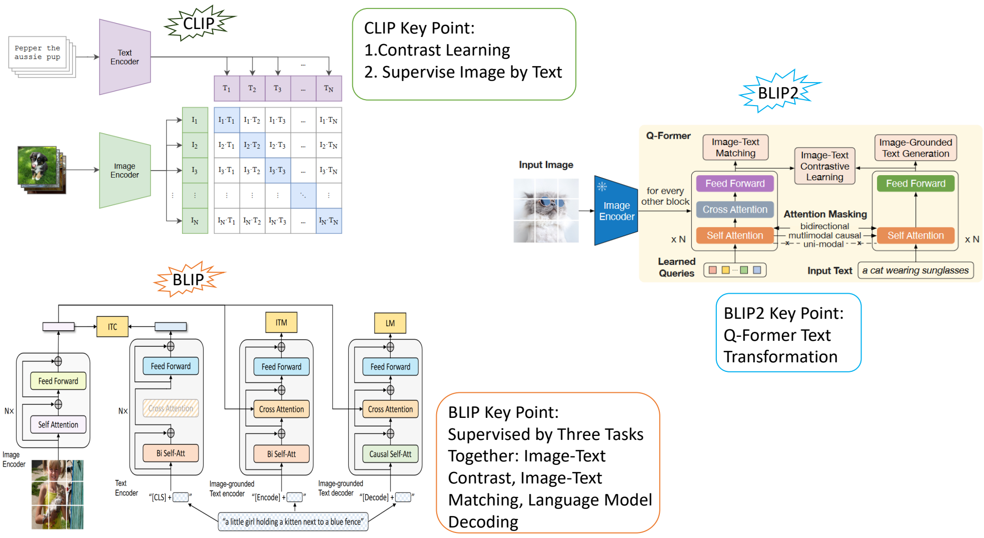
In this research, we aim to conduct this blank by utilizing pre-trained vision and language models. Our proposed approach involves guiding the language model with the vision embeddings derived from pre-trained vision or pre-trained vision-text matching models. This guidance will enable the language model to comprehend the visual content and generate textual outputs in the form of causal and temporal inferential questions.

2.2.2 Pre-trained Language Model

In recent years, pre-trained language models have revolutionized the field of natural language processing (NLP), showcasing remarkable performance across various tasks. For question generation tasks, both encoder-decoder and decoder-only architectures have been promising. Encoder-decoder models, like T5 [43], employ an initial encoding phase to process input data and generate context-rich representations, followed by a decoding phase to generate coherent and contextually relevant questions. On the other hand, decoder-only models, exemplified by GPT-2[44], leverage autoregressive generation, predicting each token based on previously generated ones, yielding fluent and contextually coherent questions.

These pre-trained models have shown promise in question generation tasks, including visual question generation[66][56]. However, their potential in capturing nuanced causal and temporal inference aspects in visual questions remains an open question. As suggested by the latest research in question generation task[66], we apply the encoder-decoder structure and the T5 language model as our baseline in our experiments.

2.2.3 Pre-trained Model in Vision to Text


CLIP
Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
BLIP2
Input Image
Image Encoder
Q-Former
Image-Text Matching
Image-Text Contrastive Learning
Image-Grounded Text Generation
for every other block
Feed Forward
Cross Attention
Self Attention
Attention Masking
bidirectional
mutlimodal causal
uni-modal
x N
Learned Queries
Input Text
a cat wearing sunglasses
BLIP
ITC
ITM
LM
Feed Forward
Self Attention
Cross Attention
Bi Self-Att
Causal Self-Att
N×
Image Encoder
Text Encoder
Image-grounded Text encoder
Image-grounded Text decoder
"[CLS] +
"[Encode] +
"[Decode] +
"a little girl holding a kitten next to a blue fence"

CLIP Key Point:
1.Contrast Learning
2. Supervise Image by Text

BLIP2 Key Point:
Q-Former Text Transformation

BLIP Key Point:
Supervised by Three Tasks Together: Image-Text Contrast, Image-Text Matching, Language Model Decoding

Figure 2.1: Difference of Vision-Text Pre-trained Models

In recent years, the advancement and robustness of vision-text matching pre-trained models have opened new avenues for research in utilizing their existing knowledge to guide language models in generating text output. These models offer a promising approach to significantly reduce the time and computation required compared to traditional pre-trained methods. Among these models, the CLIP model [43] stands out as a pioneer in leveraging cross-modal supervision to learn the matching knowledge between text descriptions and images. Subsequently, the BLIP model and BLIP2 model present a series of bootstrapping Language-Image pre-trained models. We briefly introduce their difference in this section since we will apply all of them in the methodology chapter 4. Their differences are also presented in Figure 2.1.

1. CLIP (Contrastive Language-Image Pre-training) [43]: CLIP learns to associate images and their descriptive captions in a shared embedding space, It creates a new cross-supervision method, utilizes the text embedding and image embedding cos similarity and predicts their matching degree. By pre-training on large-scale image-text datasets with 400 million image-text pairs and a simple linear projection mapping, CLIP achieves impressive results in various visual-text generative tasks, including image captioning, visual question answering, and visual storytelling. This making it a strong candidate for enhancing question generation in our context.

2. BLIP (Bootstrapping Language-Image Pre-training): BLIP [30] builds upon the success of CLIP and introduces bootstrapping strategies to improve the quality of the learned vision-text representations within their dataset. It leverages iterative bootstrapping to enhance the alignment between images and their associated textual descriptions. In addition, they propose a multimodal mixture structure, which could operate in three functionalities: unimodal encoder, image-grounded text encoder and image-ground text decoder, unifying both pre-trained classification, matching and generative tasks objectives. Both processes enhance the overall performance of BLIP in generative tasks.

3. BLIP2 (Bootstrapping Language-Image Pre-training 2.0): BLIP2 [29] further refines the bootstrapping approach introduced in BLIP. It effectively bridges the gap between vision and language. In addition, they utilise modality matching using a Q-Former pre-trained in two stages: the representation learning stage and the generative learning stage. BLIP2 demonstrates significant improvements over its predecessor and holds promise for boosting the quality of generated questions.

Our research endeavours to explore and compare the capabilities of these vision-text matching pre-trained models, in capturing the causal and temporal relationships of various events or objects within a video context. By investigating their performance in generating questions related to causal and temporal inference, we aim to identify the most suitable model for our specific research problem.

Chapter 3

Task Definition

In this chapter, We elaborate on the datasets we have utilized for our research and provide detailed explanations of their specific task attributes in Section 3.1, along with a formal definition of the VQG task. In Section 3.2, considering the limitations of existing metrics in capturing the quality of predicted questions, especially concerning their temporal and causal inference attributes, we propose a new evaluation perspective to gain deeper insights and achieve a new grounding evaluation.

3.1 Dataset and Task

The rapid development of NLP datasets especially for QA tasks in the past few years could be compared to Cambrian Explosion, with more than 80 new datasets appearing[48]. However, focusing on the video question-answering task, most datasets like MSRVTT-QA[65] and TGIF-QA[16] refer to factoid questions asking for counts, binary decision, and behaviour [1], whose answers are directly derived from the visual information. In contrast, our research focuses on a more intricate and challenging aspect of video question generation, which revolves around temporal and causal inference involving various events or objects within a video.

To address the lack of a dedicated dataset specifically tailored for visual question generation that incorporates causal and temporal inference, we turn to existing datasets designed for similar tasks, such as visual question answering (VQA). Among these, the NExT-QA dataset [60] stands out as a highly suitable choice for our Visual Question Generation (VQG) task. The NExT-QA dataset comprises both multiple-choice and

[1] Factoid question: The questions are directly asked about the visual fact according to the visual information.

open-ended questions, with a significant focus on causal and temporal inference, making it well-aligned with our research objectives.

The NExT-QA dataset encompasses a diverse range of questions, with 48% of them centred around causal inference, 29% on temporal inference, and the remaining 23% on descriptive questions. Figure 3.1 provides illustrative examples from the NExT-QA dataset, showcasing the types of questions and video content present in this dataset. By leveraging NExT-QA as our baseline dataset, we aim to explore and enhance the capabilities of generating visual questions with a focus on temporal and causal inference.

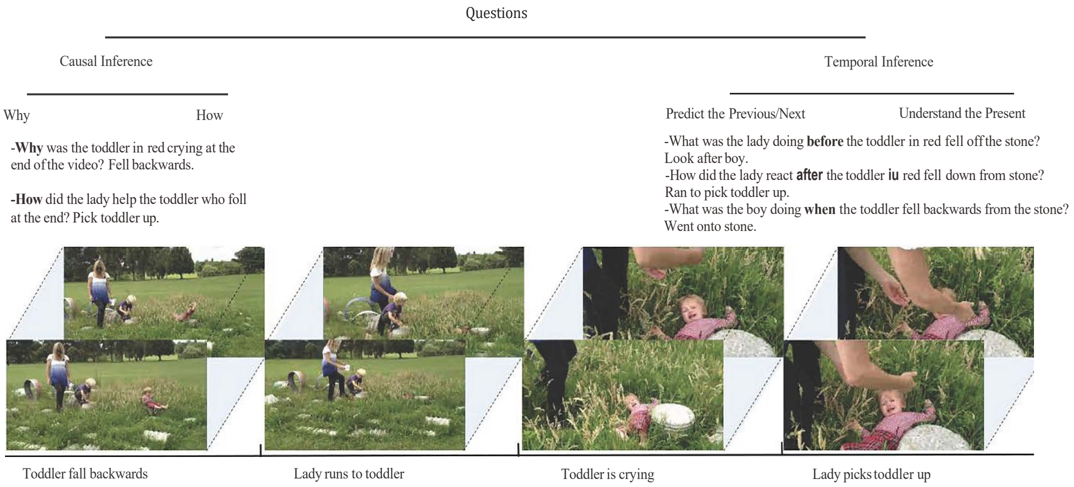

Questions
Causal Inference
Temporal Inference
Why
How
Predict the Previous/Next
Understand the Present
-Why was the toddler in red crying at the end of the video? Fell backwards.
-How did the lady help the toddler who foll at the end? Pick toddler up.
-What was the lady doing before the toddler in red fell off the stone? Look after boy.
-How did the lady react after the toddler iu red fell down from stone? Ran to pick toddler up.
-What was the boy doing when the toddler fell backwards from the stone? Went onto stone.
Toddler fall backwards
Lady runs to toddler
Toddler is crying
Lady picks toddler up

Figure 3.1: Examples in NExT-QA benchmark.

Generally speaking, an inferential visual question generation dataset comprises text format examples, each containing a question **Q** related to either temporal or causal inference, a corresponding ground truth answer **T**, and a video format example **V**. In order to simplify the task, we consider an input consisting of visual information **V**, which can be in the form of multiple images or a video sequence, along with auxiliary text information **A** representing the associated answers related to the visual content. Given this input, the core objective is to conduct research on innovative approaches to automatically generate a question **Q** that unveils the underlying causal inference or temporal inference between various events depicted in the provided visual information.

Our work mainly focuses on (1) exploring the potential of leveraging a pre-trained language model to recognize visual information (2) utilizing the frame difference and the text guidance to recognize their temporal and causal relation and generate corresponding questions. Details will be described in Chapter 4.

3.2 Evaluation Metrics

3.2.1 General Evaluation Metrics in Question Generation

Evaluating visual question generation (VQG) systems is crucial to gauge their performance. However, current VQG models mostly rely on standard language generation metrics designed for machine translation assessment. These metrics mainly gauge accuracy and similarity with reference translations, but overlook crucial aspects like inference, logic, and consistency. Commonly used metrics include:

1. BLEU (Bilingual Evaluation Understudy)[40]: Measures similarity with high-quality reference translations using n-gram matching, but misses overall syntax.
2. METEOR (Metric for Evaluation of Translation with Explicit Ordering)[2]: Similar to BLEU, it considers stemming and synonymy, using unigram precision and recall.
3. ROUGE (Recall-Oriented Understudy for Gisting Evaluation)[33]: ROUGE compares overlapping units in summaries to human-written references.
4. BLEURT (Bilingual Evaluation Understudy for Rewarding Transformers)[50]: BLEURT provides a human-like assessment by comparing generated text with references.
5. CIDEr (Consensus-based Image Description Evaluation)[55]: CIDEr evaluates image captions based on consensus among human annotations.

However, since VQG differs from machine translation, these metrics may not suit VQG evaluation. Additionally, while some metrics like CIDEr can assess the quality of ground-truth visually generated questions, they may not adequately measure inferential and reasoning questions. To address this, we aim to develop new metrics specifically focusing on inference and reasoning in VQG. These metrics will help to evaluate the grounding quality of generated questions, particularly those involving causal inference and temporal relationships, directly giving an insight into question quality and enhancing the meaningful assessment of VQG systems.

3.2.2 Overlap Grounding Evaluation Metrics

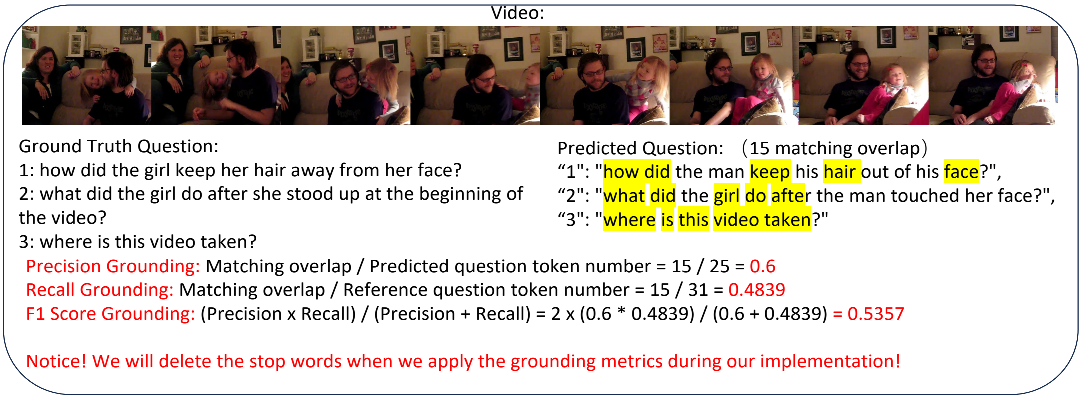
The existing evaluation metrics fail to fully capture the quality of predicted questions, especially concerning their temporal and causal inference attributes. To address this limitation, we propose a novel evaluation perspective that provides deeper insights into the grounding quality of generated questions. Our approach involves analyzing the overlap between words considering both precision and recall. We define the formula of the grounding metrics as follows:

$$\begin{aligned} Precision\ Grounding &= N_{matching\ overlap}/N_{predicted\ question\ tokens} \\ Recall\ Grounding &= N_{matching\ overlap}/N_{Ground\ Truth\ question\ tokens} \\ F1-score\ Grounding &= \frac{2*Precision\ Grounding*Recall\ Grounding}{Precision\ Grounding+Recall\ Grounding}, \end{aligned} \tag{3.1}$$

Where $N_{matchingoverlap}$ 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.

Unlike traditional metrics, these grounding metrics directly illuminate how well the predicted questions encapsulate visual content from videos. By gauging word overlaps, we effectively gauge a model's ability to comprehend visual information and contextual cues, yielding an extra layer of question quality assessment. This novel evaluation perspective enriches our grasp of visual question generation system performance, especially in generating pertinent questions related to temporal and causal inference.

To ensure the significance of information and exclude trivialities, we concentrate on word overlap, particularly nouns and verbs. By eliminating stop words like prepositions and conjunctions, we assess content-bearing words. We then measure word overlap between predicted and ground-truth questions, considering the occurrence of the same word multiple times in predictions without considering orders. These guard against bias from word repetition in predictions. Illustrative examples are shown in Figure 3.2.



Video:

Ground Truth Question:
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?

Predicted Question: （15 matching overlap）
"1": "how did the man keep his hair out of his face?",
"2": "what did the girl do after the man touched her face?",
"3": "where is this video taken?"

Precision Grounding: Matching overlap / Predicted question token number = 15 / 25 = 0.6
Recall Grounding: Matching overlap / Reference question token number = 15 / 31 = 0.4839
F1 Score Grounding: (Precision x Recall) / (Precision + Recall) = 2 x (0.6 * 0.4839) / (0.6 + 0.4839) = 0.5357

Notice! We will delete the stop words when we apply the grounding metrics during our implementation!

Figure 3.2: Examples in calculating grounding metrics

Chapter 4

Methodology

This chapter provides details of our proposed Visual Question Generation (VQG) system, with a focus on temporal and causal inference. We introduce three baseline models in Section 4.1 to establish a foundation. In Section 4.2 and 4.3, we outline the selection process of core components for our VQG system, In Section 4.4, we illustrate how we leverage visual information to guide the pre-trained language model. The results and experiments, along with their analysis and discussions, will be presented in Chapter 5.

4.1 Baseline Models

In our system, we establish baseline models for a fair comparison with the NExT-QA datasets. We begin with the Heterogeneous Graph Attention (HGA) model [17], which employs GRUs as encoders and decoders with cross attention between visual and language information. Recognizing the power of transformer-based models [54] in language generation, and the impressive results achieved by pre-trained language models in question generation tasks[66][56], we introduce another transformer-based model and a powerful pre-trained language model with text-only input as additional baselines. We aim to facilitate a straightforward comparison of various visual encoders and methods for guiding pre-trained models in generating inferential questions.

4.1.1 HGA Model

The HGA model is a prominent baseline in the NExT-QA dataset [60] for video multiple-question answering and open-end question-answering, introduced by Jiang et al.[17].

Since video shots exhibit more expressive motion compared to frame-level data, the HGA model leverages both 3D motion vectors and 2D appearance vectors to capture the richer motion expression ability of video shots compared to frame-level data. To align with the dataset benchmark, the motion vectors are abstracted by ResNet[12] and the appearance vectors are derived from ResNeXt-101[63]. Specifically, each single video is divided into N equal length clips $C = (C_1, C_2, ..., C_N)$. Each clip C_i of length $T = L/N$ is represented by 2D appearance features $V_i = \{V_{i,j} | V_{i,j} \in \mathbb{R}^{2048}\}_{j=1}^{T}$ at frame level and 3D motion features $f_i \in \mathbb{R}^{2048}$ at clip level. Importantly, the parameters of ResNet and ResNeXt-101 are frozen, which means it parameters will not be updated during training. For text input, pre-trained GloVe word embeddings are used to encode the words into embedding vectors. Both visual and language embeddings are encoded separately to obtain contextual representations, which are then fused and aligned using a designed cross-attention mechanism. A last decoder with GRUs fuses global representations of visual and language embeddings to generate output texts or classify the correct answer.

4.1.2 Pre-trained Language Model with Text-only Input

Both encoder-decoder and decoder-only structures in pre-trained language models have shown promise in question generation tasks, including visual question generation[66][56]. However, as suggested by the latest research in multiple area tasks between vision and language, especially in visual question generation [66][6], encoder-decoder structures perform better than decoder-only structures. Therefore, we apply one of the typical and powerful encoder-decoder language models —T5 [43] into our experiments. We set the text-only, with only auxiliary text as the model input as the baselines to inspect if the performance of the pre-trained language model could be improved with vision input afterwards. In other words, we would like to check if the language model could recognize the vision content given a video.

4.2 Model Components Selection

To provide readers with a clear understanding of the framework employed, we draw inspiration from the work of Zhong et al.[70] and define a common architecture for VQG, comprising four essential components: a visual encoder, an auxiliary text encoder, cross-model interaction, and an output question decoder, as depicted in Fig. 4.1. The visual encoder plays a role in processing raw videos and extracting meaningful features.

It is responsible for jointly capturing frame appearance and clip motion features. The auxiliary text encoder handles the textual information related to the visual content. Commonly used encoders include GloVe [42] and language model based embedding such as BERT [8]. To enable effective interaction between visual and textual modalities, sequential models like Transformer [54] can be employed to process the sequential data of vision and language. Finally, the question decoder is responsible for generating the output question based on the integrated visual and textual representations. To set a broad baseline for the temporal and causal inferential visual question generation task, we evaluate popular vision-text pre-trained models such as CLIP [43], BLIP [30], and BLIP2 [29]. We also compare their performance with the appearance vectors and motion vectors extracted from ResNet and ResNeXt-101, two widely used 2D and 3D neural networks, respectively.

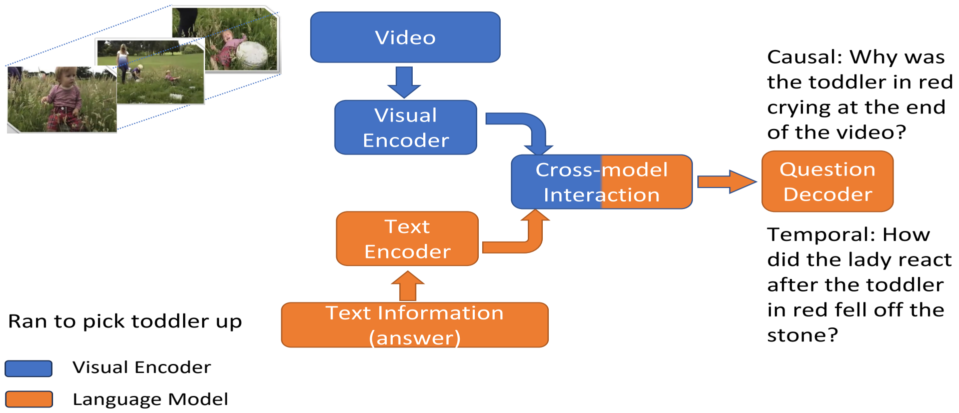

Video
Visual Encoder
Cross-model Interaction
Question Decoder
Causal: Why was the toddler in red crying at the end of the video?
Text Encoder
Temporal: How did the lady react after the toddler in red fell off the stone?
Ran to pick toddler up
Text Information (answer)
Visual Encoder
Language Model

Figure 4.1: Common Architecture Within Visual Question Generation

4.2.1 Video Encoder

In the field of visual question generation for temporal and causal inference, the NExT-QA datasets have emerged as a crucial benchmark for the video multiple-choice question-answering task. When it comes to encoding videos or frames for this task, appearance vectors and motion vectors extracted by 2D and 3D convolutional neural networks have been widely used and proven effective [61] [24] [36]. However, it is essential to recognize that these approaches primarily cater to classification tasks, which could not adapt generative tasks[66], such as question generation.

To address the limitations of the traditional 2D and 3D convolutional neural networks and exploit the potential of generative models, we turn our attention to pre-trained vision-text matching models. These models have shown exceptional performance in generative tasks[9], making them promising candidates for enhancing the visual question generation process. Unlike single 2D or 3D convolutional networks, pre-trained vision-text matching models explicitly consider the matching relationship between vision and language, allowing for more contextually relevant and coherent question generation.

In this section, we present three types of pre-trained vision-text matching models and conduct a comprehensive performance comparison with the 2D and 3D convolutional neural networks. These pre-trained models leverage large-scale datasets to learn cross-modal representations, enabling them to effectively bridge the gap between visual and textual information. By incorporating pre-existing knowledge from diverse vision-text sources, these models are poised to outperform traditional 2D and 3D approaches in the context of visual question generation tasks. The three types of pre-trained vision-text matching models we explore are CLIP[43], BLIP[30] and BLIP2[29], whose details are described in Section 2.2.3. And the results of these experiments are presented in Section 5.3.1, where we analyze and discuss the strengths and weaknesses of each approach.

4.2.2 Language Model Size Selection

Even when utilizing a consistent model structure like an encoder-decoder, variations in the number of model parameters can yield divergent performance outcomes within the same category [45]. To comprehensively examine the influence of model size on the nuanced task of recognizing relationships between events and entities in a video, we employ two distinct sizes of the T5 model: T5 Small and T5 Large. To effectively adapt these varied model sizes, we design two distinct tuning strategies (elaborated in Section 4.3.2), drawing inspiration from the recommendations put forth by [35].

4.3 Multi-modal Fusion

Visual information and textual context are often complementary in nature. The visual content provides rich details and cues that are not entirely present in the text, and vice versa. By fusing these modalities, the resulting embedding space becomes more comprehensive, allowing the language model to leverage a wider array of information during the question-generation process. Consequently, this enhances the contextual understanding of the model, leading to more accurate and relevant questions.

In the field of visual question generation, the seamless integration of information from both visual and language modalities is of utmost importance. After selecting the visual encoder and the language model for the visual question generation system, a fundamental challenge arises: the visual vectors derived from visual encoders and the

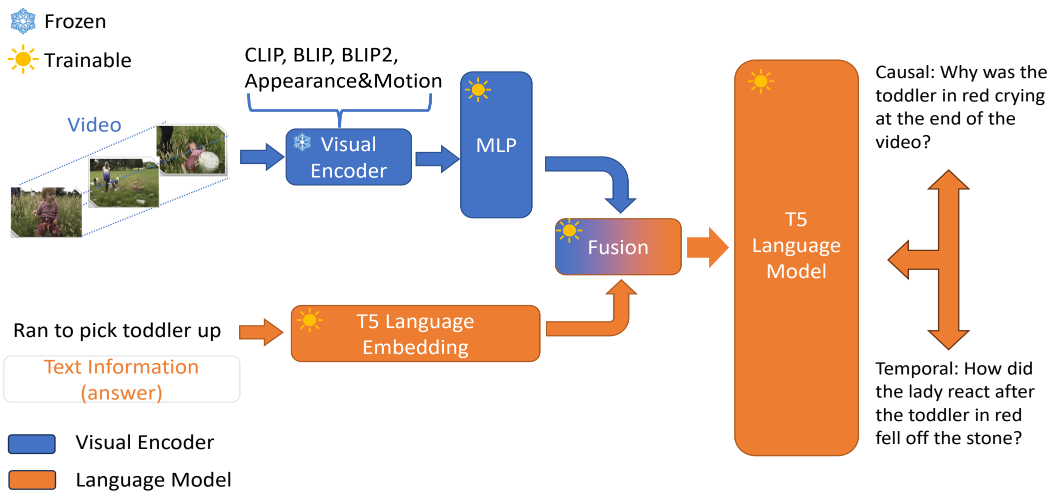

Frozen
Trainable
Video
CLIP, BLIP, BLIP2, Appearance&Motion
Visual Encoder
MLP
Fusion
T5 Language Model
Causal: Why was the toddler in red crying at the end of the video?
Temporal: How did the lady react after the toddler in red fell off the stone?
Ran to pick toddler up
Text Information (answer)
T5 Language Embedding
Visual Encoder
Language Model

Figure 4.2: Our fusion framework Within visual question generation

text embeddings from language encoders exist in separate spaces. Therefore, the core questions that demand attention are how to fuse or 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. The total frame is presented in Figure 4.2. In Section 4.3.1, we introduce one of the direct but powerful methods to connect vision and language spaces. In Section 4.3.2, we introduce another method to finetune large-size language models effectively.

4.3.1 Concatenate Vision and Language

With the popularity and recognition of language models [45][27][3] and visual transformer [9][43], multi-modal interaction has been critical for the language model to effectively recognize visual cues and context. Inspired by one of the latest methods Clipcap [37], we propose a direct but powerful technique to connect vision and language spaces effectively. By utilizing cutting-edge fusion techniques, we combine visual embeddings and language embeddings to create a unified embedding space. Specifically, given auxiliary text input words $w_V^1, w_V^2, ..., w_V^i$ 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 $x_V^1, x_V^2, ..., x_V^i$. 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:

$$p_V^1, p_V^2, ...p_V^k = F(visual\_encoder(x_V^1, ..., x_V^i)). \tag{4.1}$$

where each vector p_V^k has the same dimension as a word embedding of language models, and the choice of visual encoder is detailed in Section 4.2.1 We then concatenate the obtained visual embedding to the auxiliary input text embeddings:

$$Z_V = p_V^1, ..., p_V^k, t_V^1, ..., t_V^i. \tag{4.2}$$

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:

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{j=1}^{\ell} \log p_\theta(q_j^i | Z_V, q_1^i, ... q_{j-1}^i) \tag{4.3}$$

, where N is the number of videos, ℓ is the length of the predicted questions, p_θ is the probability of ground-truth tokens. Details of training methods will be described in Section 4.3.2

4.3.2 Two Stage Fine Tuning

The challenge of translating between representations of visual encoders and language models during training poses a fundamental hurdle in multi-modal fusion for visual question generation. Most research in multi-modal generative tasks focuses on unifying the multi-modal embedding space during pre-training [57] [18] [41], often overlooking the potential of leveraging and guiding existing pre-trained models to excel in generative tasks. This section introduces a novel two-stage fine-tuning approach to train the visual question generation system effectively, inspired by the works of [35].

Stage 1: Fine-tuning for Feature Alignment. In this initial stage of the two-stage training process, we prioritize feature alignment between the visual encoder and the language model. Drawing inspiration from the works of [31] and [37], which accommodate pre-trained models to unfamiliar tasks through learning a prefix, we adopt a similar approach. Instead of fine-tuning the entire model, we only train a parameter mapping network F (as shown in Equation 4.1) to align the video features V with the language model's word embeddings. By focusing solely on optimizing the parameter of mapping, we achieve a lightweight model while aligning the visual tokenizers.

Stage 2: Fine-tuning End-to-End. Once the training loss of Stage 1 has converged, we proceed to Stage 2, which involves fine-tuning the visual question generation system end-to-end. Drawing insights from prior works [35][37], which suggest that fine-tuning visual encoders does not significantly improve the resulting quality but introduces complexity and cost, we opt to keep the visual encoder weights frozen. In this stage, we continue to update both the pre-trained weights of the projection layer and the language model. This approach efficiently optimizes the language model's performance, particularly for large-size models, in generating temporal and causal questions.

In summary, the two-stage fine-tuning approach addresses the core challenge of multi-modal fusion by effectively aligning visual and textual information, and subsequently optimizing the language model's performance for temporal and causal inference during visual question generation.

4.4 Negative Causal and Temporal Inference Abstraction Methods

Upon amalgamating visual and linguistic data, the subsequent phase involves delving into the realm of abstracting causal and temporal inferences from various events and entities portrayed within a video. This section sheds light on a quartet of distinct methodologies, each characterized by its inclination towards unraveling the intricate tapestry of inferential connections that underlie the visual content.

4.4.1 Vision Projection Matrix Choice

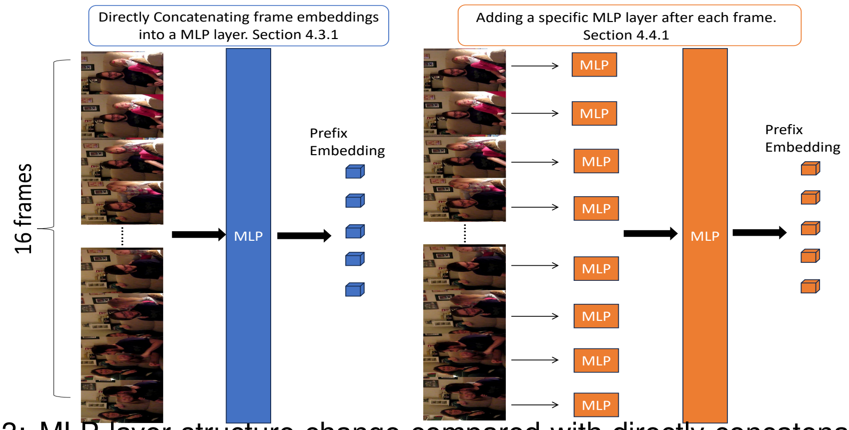

Directly Concatenating frame embeddings into a MLP layer. Section 4.3.1
Adding a specific MLP layer after each frame. Section 4.4.1
16 frames
MLP
Prefix Embedding
MLP
MLP
MLP
MLP
MLP
MLP
MLP
MLP
MLP
Prefix Embedding

Figure 4.3: MLP layer structure change compared with directly concatenating frame embedding into a large MLP layer detailed in Section 4.3.1.

An initial and intuitively straightforward approach involves the development of a meticulously detailed video projection matrix, surpassing the configuration delineated in Section 4.3.1. In contrast to the previous methodology that entailed a simple concatenation of 16 frame vision embeddings into a Multi-Layered Perceptron (MLP) projection, our endeavour here was to encapsulate the intricate nuances of each frame's characteristics. To achieve this, a distinct MLP layer with the same hidden layer of the MLP layer described in Section 4.3.1 was crafted for every individual frame, and subsequently augmented by an additional MLP tasked with projecting the 16 frame embeddings onto a linguistic embedding with a prefix length of 5, shown in Figure 4.3.

4.4.2 Contradictory Frame Comparison

In the pursuit of unravelling the intricate causal and temporal relationships within a video's array of events and entities, an instinctive avenue to explore involves harnessing the disparities between consecutive frames. This endeavour seeks to leverage frame differences to guide the language model's recognition of these variations and subsequently express them through the generated questions. Our approach involves two distinct strategies for frame comparison, both of which hinge on the CLIP vision encoder [43].

1. **Global Frame Comparison:** We abstract 16 frames at uniform intervals throughout the video's duration. These frames are transformed into vision embeddings through the CLIP encoder. Pairwise combinations of frames are formed, with the cosine similarity between their corresponding embeddings serving as a measure of their similarity. Among these frame pairs, we pinpoint the duo exhibiting the lowest cosine similarity as the most contradictory frames, capturing the divergent aspects of the video. To encapsulate this contrast, an MLP layer after the visual encoder is employed to project these two frames onto the language embedding.

2. **Local Frame Comparison:** Expanding on the global approach, our local contrast methodology undertakes a more nuanced route. Once again, we select pairs of frames and gauge their cosine similarity. But during training, 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. Armed with these contextual cues, we measure the cosine similarity between the identified relevant frame and other frames in the pair. Subsequently, the frame displaying the lowest cosine similarity with the contextually chosen frame is selected. Analogous to the global contrast approach, an MLP layer imparts the selected frame pair onto the language embedding.

4.4.3 Contrastive Learning on Unifying Vision and Language Embedding

In the pursuit of harnessing the nuanced interplay between frames within a video, a seemingly direct and effective avenue is the application of contrasting learning methods. This strategy seeks to amplify the contrast and divergence between various elements by maximizing a lower bound of mutual information between pairs of variables. In our experimental foray, we employed the infoNCE loss function [39], a widely embraced paradigm for contrastive learning [59]. 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:

$$\mathcal{L}_k = -\log(\frac{e^{f(x^+,c)}}{e^{f(x^+,c)} + \sum_{i=1}^{k} e^{f(x_i^-,c)}}) \tag{4.4}$$

In our experimental domain, building upon the frame comparison methodology, we derived positive samples from two distinct frame pairs:

1. The global contradictory frame pair assumes the role of positive embeddings. To extract these, we executed a process akin to that elucidated in Section 4.4.2. Subsequently, we designated the remaining frames, paired with the second frame from the global contradictory set, as negative samples. In the InfoNCE loss formula, x^+ signifies the positive sample language embedding, while x_i^- denotes the negative sample language embeddings. The variable c encapsulates the embedding of the second frame within the global contradictory frame pair.

2. Analogous to the process delineated in Section 4.4.2, the local contradictory frame pair was employed as the positive sample set. Correspondingly, the remaining frames were paired with the second frame from the local contradictory set, constituting the negative vision samples. Within the formula, x^+ signifies the positive sample language embedding, x_i^- represents the negative sample language embeddings, and c encapsulates the embedding of the second frame from the local contradictory pair.

The integration of the contrastive learning loss was interwoven with the pre-trained language model loss. Formally, the total loss function was defined as:

$$\mathcal{L}_{Total} = \mathcal{L}_{language\ model} + \mathcal{L}_k \tag{4.5}$$

4.4.4 Visual-Semantic Arithmetic Inferential Relation Abstraction

To harness the interplay of differences between frames within a video, a straightforward approach involves subtracting frame embeddings—an intuitive representation of the vector direction—thus capturing the inherent relationships between vectors. Recent investigations [53, 10] have unveiled the intricate taxonomy held within the CLIP multi-modal representation. Notably, their findings underscore the potential for uncovering relationships by subtracting these representations, particularly among different images. Inspired by this, we sought to adapt their CLIP loss function to augment the guidance for our language model in recognizing relationships, notably causal and temporal, between diverse frames.

Initially, we 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:

$$D_i^k \propto \exp\left(\frac{F_{cos}(E_{Text}(s_i^k), E_{frame}(I))}{\tau_c}\right), \tag{4.6}$$

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 hyperparameter 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:

$$\mathcal{L}_{CLIP} = CE(D_i, x_{i+1}). \tag{4.7}$$

This loss fosters words that yield higher CLIP matching scores between images and the generated sentences. In turn, it encourages the language model to discern the relationships between frames, encompassing causal and temporal inferences. Formally, the total loss function is defined as:

$$\mathcal{L}_{Total} = \mathcal{L}_{language\ model} + \mathcal{L}_{CLIP} \tag{4.8}$$

Chapter 5

Experiment, Results and Analysis

This chapter presents the experimental design, outcomes, and analysis of our study on visual question generation concerning temporal and causal inference questions. In Section 5.1, we outline the experimental settings. Section 5.2 introduces the experiment of the baseline models, which incorporates GRUs (Gated Recurrent Units) and Language models with text-only input. Continuing in Section 5.3, we delve into our experiments with multi-modal information direct concatenation. Within this section, we undertake a comparative analysis of different video encoders and evaluate the effects of employing various language model sizes. Next, in Section 5.4, we provide intricate details concerning the inferential methods and the selection of frames within a video input. Finally, Section 5.5 constitutes a discussion of the results acquired from our experiments.

Through these meticulously designed experiments and thorough analysis, we aim to advance the field of visual question generation, particularly in the domain of temporal and causal inference questions.

5.1 Experiment Settings

We implement all experiments with the T5 language model and vision-text matching models based on Huggingface[1] and NExT-QA framework[2]. We use Adam[20] as our optimizer and we set the learning rate as 0.0005. All experiments are based on PyTorch 2.0 and Python 3.10.

[1] https://github.com/huggingface

[2] https://github.com/doc-doc/NExT-QA

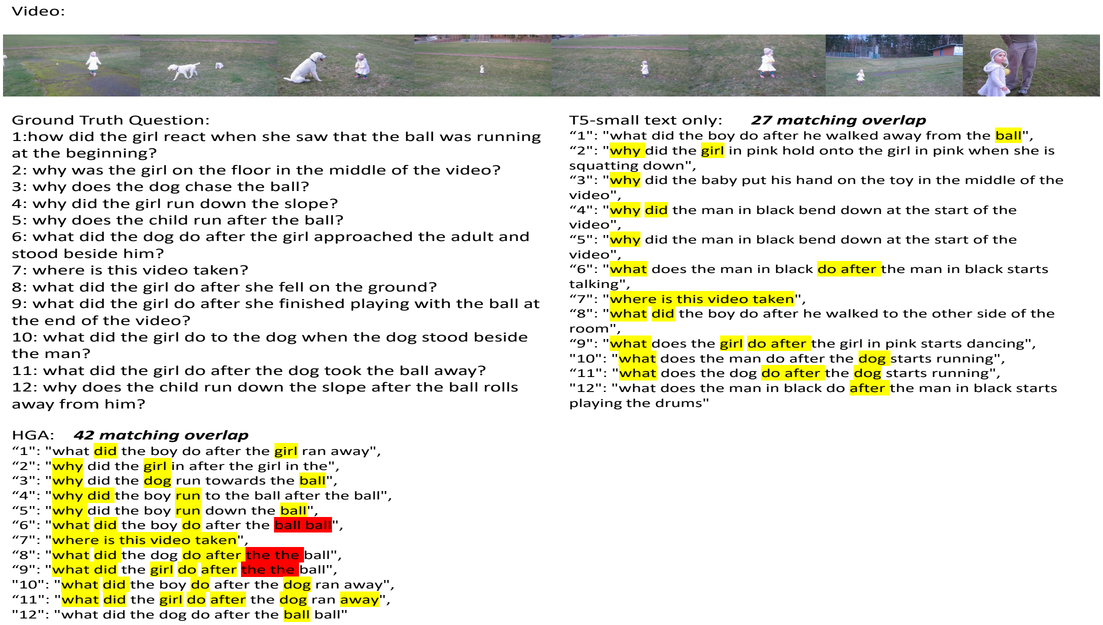
5.2 Stage 0: Baseline Model Preparation

We evaluated our baseline models following the method outlined in Section 4.1, with results summarized in Table 5.1. The HGA model [17], incorporating video input, demonstrated superior grounding proficiency across the grounding metric, while the T5 model excelled in BLEU, METEOR, BLEURT, and CIDEr, indicating better question quality.

model	B	RL	M	BL	C	Grounding
HGA[17]	0.1248	**0.4128**	0.3101	-1.1031	0.8271	**0.3248**
T5 Small Text Only	**0.1269**	0.3857	**0.3276**	-0.9986	**0.8480**	0.2957
T5 Large Text Only	0.1239	0.3851	0.3237	**-0.9808**	0.8353	0.2987

Table 5.1: Baseline Model Evaluation Performance. B is BLEU, RL is ROUGEL, M is METEOR, BL is BLEURT, C is CIDEr, Grounding is the grounding metric

The HGA model's lower question quality arises from two main factors. Firstly, it tends to generate repetitive words like "the," affecting overall quality, as illustrated in Figure 5.1's red scope. Secondly, HGA questions are shorter and less fluent than T5 questions, further contributing to this disparity (Figure 5.1). Although HGA's questions have higher matching overlap, implying better vision recognition, they lack fluency and length. Notably, the HGA model's performance closely matches T5's in BLEU, METEOR, BLEURT, and CIDEr due to its elevated precision but comparatively lower recall which is even lower than those of T5 especially in causal and temporal questions, shown in the Appendix Table A.1. While it excels at recognizing video content due to its video input, it generates fewer words than T5. Since the BLEU and BLEURT both have 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. However, since our grounding metric ignores stopwords and considers relevant words to the vision content such as nouns and verbs, precision will take an advantage in the evaluation compared to recall, thus the HGA model achieves a significant improvement compared to the T5 model. 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, especially in causal and temporal questions. This leads us to choose T5 as the foundation for subsequent experiments.



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 5.1: Baseline Performance. Yellow scopes represent the matching overlap compared with the ground truth questions. Red scopes represent the repetitive words.

5.3 Stage 1: Multi-modal Concatenation Experiment

5.3.1 Video Encoder Comparison

visual model	B	RL	M	BL	C	Grounding
None Text Only	0.1269	0.3857	0.3276	-0.9986	0.8480	0.2957
App&Mot[12][63]	0.1348	0.3958	0.3353	-0.9586	0.8816	0.3092
CLIP[43]	0.1564	**0.4216**	0.3594	**-0.8284**	1.0366	**0.3505**
BLIP[30]	0.1562	0.4179	0.3584	-0.8504	1.0205	0.3425
BLIP2[29]	**0.1583**	0.4210	**0.3599**	-0.8422	**1.0488**	0.3455
BLIP2 Q-form[29]	0.1520	0.4135	0.3537	-0.8656	0.9960	0.3353

Table 5.2: Visual encoders performance with T5 small language model following Section 4.3.1 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, BL is BLEURT, C is CIDEr, Grounding is the grounding metric.

Following the methodology described in Section 4.2.1, we assess the performance of different vision encoders, and the results are summarized in Table 5.2. Among the evaluated encoders, the CLIP[43] and BLIP2[29] models stand out, exhibiting superior performance compared to other vision encoders. The CLIP model outperforms all

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 5.2: 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.

other encoders in terms of ROUGEL, BLEURT, and grounding metrics, showcasing its proficiency in recognizing visual content and facilitating pre-trained language model guidance. Conversely, the BLIP2 model excels in BLEU, METEOR, and CIDEr, indicating its ability to generate high-quality predicted questions.

Figure 5.2 provides a visual representation of the comparison between the question generation systems using the CLIP and BLIP2 encoders. The system utilizing the BLIP2 encoder generates more detailed questions (indicated by the red scope) compared to those derived from the CLIP encoder. However, the matching overlap between the question generation system with the CLIP encoder and the ground truth questions (shown in the yellow scope) is higher than that with the BLIP2 encoder. This suggests that the CLIP encoder performs better in generating questions that closely match the vision video and the ground truth questions.

Furthermore, it is noteworthy that the evaluation of the CLIP and BLIP2 models reveals their specific strengths in different aspects. The BLIP2 model is particularly adept at visual dialogue or answering tasks, while the CLIP model excels in visual commonsense reasoning, as suggested by [43] and [29]. Taking into account the trade-off between question quality, vision content recognition, and inference reasoning, we have made the decision to employ the CLIP model as our image encoder for the subsequent experiments.

5.3.2 Language Model Size Comparison

model	B	RL	M	BL	C	Grounding
T5 Small One Stage	0.1564	0.4216	0.3594	-0.8284	1.0366	0.3505
T5 Small Two Stage	0.1559	0.4181	0.3594	-0.8409	1.002	0.3453
T5 Large One Stage	0.1459	0.4025	0.3459	-0.9046	0.9449	0.3249
T5 Large Two Stage	**0.1572**	**0.4281**	**0.3634**	**-0.8000**	**1.0657**	**0.3573**

Table 5.3: 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, BL is BLEURT, C is CIDEr, Grounding is the grounding metric.

This section investigates language model sizes for the T5 pre-trained model in video question generation. Following the approach in Section 4.3.2, we evaluate T5's performance across various sizes, presenting results in Table 5.3. T5 large outperforms T5 small as expected due to its larger parameter count, effectively storing a more extensive repository of linguistic knowledge compared to T5 small[45].

A notable insight emerges through two-stage tuning. T5 large with two-stage tuning improves over one-stage tuning, while T5 small falters. Our observations yield two primary findings:

1. While the T5-small model demonstrates inferior overall performance with the two-stage tuning compared to the one-stage method, the consistent application of two-stage tuning notably enhances token-level matching overlap across word types such as nouns and verbs. This improvement holds true regardless of the T5 language model's size, as detailed in Appendix A.2. This underscores the enhanced visual content recognition ability of the T5 model through the two-stage tuning methodology. As proposed by [35], we argue that this improvement can be attributed to the initial stage's weight initialization and warming-up of the projection matrix. This process facilitates better alignment between the projection matrix's weights and the pre-trained language model's weights, leading to more

effective fine-tuning, as opposed to directly fine-tuning the projection matrix's weights.

2. T5 small with two-stage tuning generates more repetitive questions than one-stage, contrasting T5 large where repetition decreases. An illustrative example is presented in Appendix Figure A.3. Although the total performance of different sizes of the T5 models is close, focusing on causal and temporal questions, we find that the T5 large model has a higher performance with nearly 2%-3% than that of the T5 small model on causal questions but achieves a close performance on temporal questions, shown in Append Table A.3. This reveals the potential of our tuning method in guiding the language model to recognize the causal relationship between events and entities and a future direction could research how to guide the temporal relationship.

Drawing upon the insights proposed by [35] and [45], we argue that the second observation in repetitive questions is largely due to the disparity in model parameters between T5 small and T5 large. The limited parameter capacity of T5 small constrains its ability to memorize and learn the nuances of generating diverse questions while accommodating similar video frame inputs and answers. In contrast, the expanded parameter space of T5 large enables a deeper comprehension of inputs, greater generalization capabilities, and subsequently, a reduction in the generation of redundant and repetitive questions.

5.4 Stage 2: Causal and Temporal Inference Abstraction

In the forthcoming section, we present the outcomes of our diverse 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. Despite the absence of the desired performance outcomes, our analysis serves as a valuable exploration, offering glimpses into the intricate interplay of causal and temporal dynamics within video content.

5.4.1 Vision Projection Matrix Comparison

Employing the methodology elucidated in Section 4.4.1, this section delves into the assessment of various projection matrix techniques. The culmination of our efforts is distilled in Table 5.4.

model	B	RL	M	BL	C	Grounding
Video MLP	**0.1564**	**0.4216**	**0.3594**	**-0.8284**	**1.0366**	**0.3505**
Video 16to5 MLP	0.1549	0.4170	0.3574	-0.9323	0.9722	0.3415

Table 5.4: Vision Projection Matrix Performance. Both experiments are conducted with CLIP image encoder and T5 small pre-trained language model. 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, BL is BLEURT, C is CIDEr, Grounding is the grounding metric

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?

Video 16to5 MLP: (***43 matching overlap***)
9": "what is the man wearing on his head",
"1": "what does the man do after the girl puts her hand on the man s face",
"10": "where is this happening",
"0": "how does the man use body language to express what he wants to say",
"6": "why did the man in black hold the girl s hand",
"4": "what does the man do after the girl sits down",
"8": "what does the man do after the girl turns to the left",
"3": "why did the man in black bend down at the end of the video", "2": "why did the man point to the sofa after the girl walked away", "12": "what does the man do after the girl turns to the left",
"11": "how did the man ensure he can see clearly",
"7": "why did the man laugh when the girl is talking",
"5": "why did the man move his arms backwards in the middle of the video"

Video MLP: (***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 5.3: Vision Projection Matrix Performance. Yellow scopes represent matching overlap with ground truth questions.

Contrary to our initial expectations, a noteworthy trend emerged from our results. Specifically, the methods involving the direct concatenation of vision embeddings from the CLIP image encoder to the language embedding's prefix outperformed those that employed the addition of MLP layers to each frame before concatenating with the language embedding, including grounding metrics on causal and temporal questions(Appendix Table A.4). A concrete instance illustrating this divergence is portrayed in Figure 5.3, where a significant discrepancy in the number of matching overlaps between the ”Video 16to5 MLP” method and the ”Video MLP” method is evident.

This finding carries an implication: the blind proliferation of MLP layers, even when applied to individual frames, fails to capture the fine-grained details and inferential relationships of visual content. Consequently, this approach falls short in guiding the language model to generate inferential questions that accurately reflect the subtle causal and temporal relationships embedded within the video.

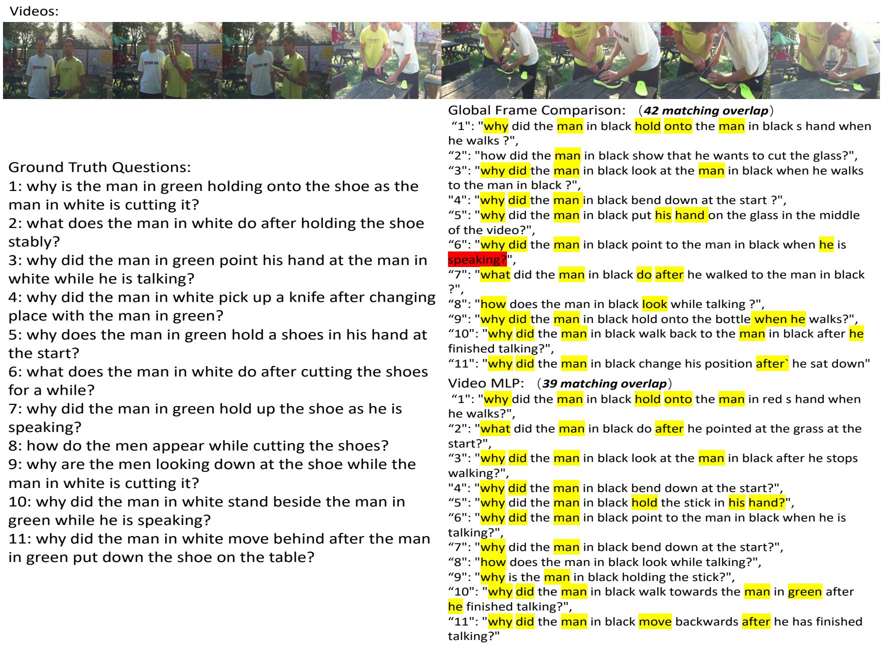
5.4.2 Frame Comparison Based on CLIP

model	B	RL	M	BL	C	Grounding
All 16 frames(Video MLP)	**0.1564**	**0.4216**	**0.3594**	**-0.8284**	**1.0366**	**0.3505**
Two frames(Random Select)	0.0796	0.3128	0.2173	-1.2445	0.2520	0.2082
Two frames (Global Frame Comparison)	0.1538	0.4165	0.3578	-0.8422	1.007	0.3417
Two frames (Local Frame Comparison)	0.1315	0.3946	0.3316	-0.9386	0.8576	0.3095
Two frames(Fixed Selection) Frame 1&16	0.1526	0.4161	0.3549	-0.8679	0.9745	0.3407

Table 5.5: 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, BL is BLEURT, C is CIDEr, Grounding is the grounding metric.

In this section, we apply the approach detailed in Section 4.4.2 to assess two distinct methods of frame comparison. The summarized outcomes are presented in Table 5.5, yielding several noteworthy findings:

1. The global frame comparison method, while slightly behind direct vision embedding concatenation as a language prefix, outperforms the baseline, especially for causal and temporal questions with a substantial 20% boost in temporal questions. It demonstrates competitive performance in causal questions and over 1.5% higher temporal performance in grounding metrics compared to direct vision concatenation (Appendix Table A.6).These findings underscore the efficacy of the global frame comparison method in abstracting causal and temporal relationships within input videos, effectively guiding the T5 language model toward generating insightful inferential questions. Importantly, the global frame method, with only 73M parameters, surpasses the direct concatenation approach (135M), still maintaining competitive performance. Notably, the global frame method also displays proficiency in recognizing verbs and pronouns in comparison to the direct concatenation techniques. For specific examples and instances of matching



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 5.4: 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.

overlaps across different categories, please refer to Appendix A.5, along with an illustrative example showcased in Figure 5.4.

2. In contrast, the local frame comparison method yields inferior results compared to its global counterpart across all evaluation metrics. 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 effectively deduce relationships between events and entities within videos.** The method of random selection introduces the highest level of inconsistency 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 reveals certain limitations. While instances of accurate frame selection aligned with questions and answers are observed, inherent challenges persist: (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 [60] 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. Detailed examples highlighting these challenges are provided in Appendix Figure A.1. These issues exacerbate inconsistencies and disorderliness in input frames between training and inference, resulting in comparatively poorer performance compared to the global frame comparison method. Significantly, the global frame method introduces the least inconsistency, consistently measuring cosine similarity and selecting the least similar frame pair for language model input. These inherent contradictions effectively mitigate the degree of disparate frame relationships.

3. To further corroborate our argument, we conduct an additional experiment where the initial and final frames are consistently selected as the video input for the language model, as outlined in the fifth row of Table 5.5. 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 closely emulate consistent relationships to enhance frame-based techniques.

5.4.3 Contrastive Learning Based on Frame Comparison

model	B	RL	M	BL	C	Grounding
Global Frame Comparison baseline	0.1538	**0.4165**	0.3578	**-0.8422**	1.007	0.3417
Global Frame Comparison Contrast	**0.1555**	0.4164	**0.3601**	-0.8767	**1.010**	0.3383
Local Frame Comparison Contrast	0.1531	0.4165	0.3555	-0.9456	1.001	**0.3426**

Table 5.6: Contrasting Learning Performance Based on Global Frame Comparison. B is BLEU, RL is ROUGEL, M is METEOR, BL is BLEURT, C is CIDEr, Grounding is the grounding metric.

This section employs the approaches from Section 4.4.3 to evaluate two contrasting learning methods, both rooted in global frame comparisons as discussed in Section 4.4.2. Our analysis, in the context of the global frame comparison baseline outlined in

Section 5.4.2, is summarized in Table 5.6.

Out of our expectations, both global frame contrast and local frame contrast methods outperform the baseline in specific metrics, yet their overall performance remains closely comparable, differing by less than 0.01, except for BLEURT. This prompts us to question their utility. We delve deeper, inspecting causal and temporal question outputs, and calculating overlap across distinct word categories. More details are shown in Appendix Table A.7 and Appendix Table A.8, with an example in Appendix Figure A.4. Despite the marginal disparity in overall performance between the baseline and the two contrasting learning methods, a closer inspection reveals that contrasting learning can facilitate the language model's ability to discern nuanced details within the video, such as characters, colours, verbs, and tense, as illustrated within Appendix Figure A.4's red scope. The local contrast method also improves temporal question grounding metrics by 1-2% over the global frame comparison baseline. These discoveries underscore the potential of applying contrastive learning to bridge the gap between the visual and language embedding spaces. It augments the language model's capacity to comprehend video content and subsequently generate inferential questions especially temporal relationships.

Exploring the rationale behind the observed similarity in performance between contrasting learning and the baseline, a perspective shared by [59] and [45], we posit the following considerations: (1) The negative sample pool in our methods is relatively limited, constraining the model's ability to discern mutual information between positive and negative samples. Given the video's continuous nature, some nearly abstracted frames exhibit visual similarity, further complicating the model's differentiation process. (2) The parameters of the T5 small model are inherently constrained, limiting its capacity to encompass the entirety of knowledge necessary for recognizing all video events and entities, as well as grasping the subtleties distinguishing positive and negative samples during contrastive learning.

5.4.4 Visual-Semantic Arithmetic Inferential Relation

The summarized findings are outlined in Table 5.7. We observe that the performance of the visual-semantic arithmetic method closely resembles that of the baseline approach, directly concatenating vision embeddings. This suggests that supplementing the visual-semantic arithmetic with CLIP loss may not yield significant improvements.

To further validate the potential of the visual-semantic arithmetic method, we com-

model	B	RL	M	BL	C	Grounding
Video MLP	0.1564	**0.4216**	0.3594	**-0.8284**	**1.0366**	**0.3505**
CLIPloss top word 100	**0.1568**	0.4184	**0.3602**	-0.8295	1.0359	0.3460

Table 5.7: 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, BL is BLEURT, C is CIDEr, Grounding is the grounding metric.

pare the questions generated by the two frame selection techniques and scrutinize whether their disparities are accurately portrayed in the generated questions. Specific examples of successes and shortcomings are presented in Appendix Figure A.2. Additionally, we examine the generated questions in causal and temporal types and compare their matching overlap levels with the baseline. Performance details of causal and temporal questions are shown in the Appendix Table A.9 and a concrete instance is illustrated in Figure 5.5. It is found that the visual-semantic arithmetic method outperforms temporal questions with a 1-2% increase compared with the direct vision concatenation. It's apparent that the visual-semantic arithmetic method exhibits a higher degree of matching overlap compared to the baseline. Notably, the method adeptly recognizes time adverbs (e.g., "when"), underscoring its potential to discern temporal details and providing support for the increase of its performance on temporal questions.

Drawing insights from the examples presented in the Appendix Figure A.2, the method's effectiveness, and the CLIP model's subtraction semantic attribute as suggested by [53] in alignment with [43], we argue that the multi-model concatenation methods may fall short in enabling the language model to comprehensively discern the complete spectrum of visual relationships among the most contrasting frame pair within a video.

5.5 Discussion

In this thesis, we introduce an inferential framework for generating causal and temporal questions based on videos and auxiliary text. The Stage 1 experiments highlight the efficacy of our visual concatenation technique in enhancing performance compared to the baseline HGA [17] and the T5 model [45] using text-only input. In Stage 2,

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 5.5: 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.

although our causal and temporal inference methods do not surpass the Stage 1 baseline, they lay the groundwork for future research avenues while showcasing their potential.

The Vision Projection Matrix Comparison underscores that an excessive proliferation of MLP layers fails to capture the subtle nuances within input visual content. The Frame Comparison analysis emphasizes the critical role of establishing consistent relationships between input frames during both training and inference. This consistency is vital for enabling the language model to adeptly infer relationships among events and entities within videos. Our exploration of the Contrastive Learning method illuminates the limitations tied to the number of negative sample pools and language model size. Promising research avenues could involve expanding negative samples by random frame selection from other videos and considering more parameter-rich models like the T5 large model. Lastly, the Visual-Semantic Arithmetic method underscores a promising research direction - guiding pre-trained language models to recognize arithmetic relationships, such as subtraction and addition, among diverse frames.

In summary, our research contributes a framework to video question generation, particularly about causal and temporal inference. While some methods exhibited comparable performance to baselines, they unveil intriguing avenues for future exploration.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This thesis addresses the research gap in aligning machine-generated visual questions with human cognitive processes, specifically in the domain of video visual question generation (VQG). Instead of focusing on simple factual queries, our study delves into generating inferential questions that involve causal and temporal inference. While prevailing pre-trained model methods have demonstrated excellence, they come with high computational demands, and the potential of leveraging relationships among diverse pre-trained models in multi-modal learning remains untapped. To bridge these gaps, we propose an innovative framework that employs vision-text matching pre-trained models to facilitate pre-trained language models in identifying event-entity relationships within videos and generating inferential questions.

Our video VQG framework comprises four key components: visual encoder, text encoder, cross-modal interaction, and question decoder. To establish a robust performance for our framework, we conduct a comparison of four distinct visual encoders and two sizes of pre-trained language models, coupled with a specific training approach. Recognizing the limitations of existing evaluation metrics in the VQG realm, we introduce a grounding metric to provide direct insights into the language model's ability to comprehend visual content, thereby enhancing evaluation. Moreover, we propose a direct and potent method for integrating vision and language information. Lastly, we present four diverse techniques, encompassing projection layer design, frame comparison, contrastive learning, and visual-semantic arithmetic. These methods enhance the abstraction of causal and temporal relationships within videos, guiding the language model towards superior inferential question generation.

Our experimental results underscore the efficacy of our proposed video VQG framework. Notably, a substantial enhancement of approximately 3-5% across all evaluation metrics, achieved through the utilization of visual encoders with text-only inputs, underscores the effectiveness of our framework in promoting visual content recognition by the language model. We undertake an array of experiments to compare visual encoders and language model sizes, pinpointing the most effective configurations for subsequent inferential experiments. Furthermore, while our advanced abstraction methods for causal and temporal relationships yield comparable outcomes to direct concatenation of vision and language embeddings, they provide referential research in MLP layer design. Additionally, our experiment results suggest promising directions for future exploration, including ensuring consistency in frame selection, augmenting negative samples for contrastive learning, and guiding pre-trained language models to recognize arithmetic relationships within image pairs.

6.2 Future Work

While our innovative framework successfully generates inferential questions related to causal and temporal inference, our experiments and analyses illuminate avenues for future advancement. Firstly, our frame comparison analysis underscores the importance of consistent relationships between input frames during both training and inference. Investigating methods to enhance frame consistency holds promise.

Secondly, our study of the contrastive learning method highlights limitations tied to the number of negative samples and the size of the language model. Future research could consider augmenting the pool of negative samples by substituting question-unrelated frames with frames from different videos, thus enhancing diversity. Additionally, increasing the language model size, such as transitioning from advanced T5-small to T5-3B, holds potential, as larger language models exhibit stronger capabilities in comprehending frame distinctions.

Finally, the visual-semantic arithmetic method paves the way for a promising research avenue. Given that the vision-text matching pre-trained model can grasp frame differences through embedding subtraction and addition, while the pre-trained language model cannot, a subsequent exploration could focus on conveying frame differences from vision-text matching pre-trained models to pre-trained language models. This research direction aims to guide pre-trained language models in recognizing arithmetic relationships among diverse frames.

Bibliography

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] Satanjeev Banerjee and Alon Lavie. 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, June 2005. Association for Computational Linguistics.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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

[5] David Chen and William Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[6] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021.

[7] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

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

[11] Zhaoyu Guo, Zhou Zhao, Weike Jin, Zhicheng Wei, Min Yang, Nannan Wang, and Nicholas Jing Yuan. Multi-turn video question generation via reinforced multi-choice attention network. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5):1697–1710, 2020.

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

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[14] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1233–1239, 2016.

[15] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

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

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

[18] Zaid Khan, Vijay Kumar BG, Samuel Schulter, Xiang Yu, Yun Fu, and Manmohan Chandraker. 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, 2023.

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

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[21] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

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

[23] Geoffrey T LaFlair, Andrew Runge, Yigal Attali, Yena Park, Jacqueline Church, and Sarah Goodwin. Interactive listening–the duolingo english test. Technical report, Duolingo Research Report DRR-23-01). https://go. duolingo. com/interactive . . . , 2023.

[24] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9972–9981, 2020.

[25] Che-Hao Lee, Tzu-Yu Chen, Liang-Pu Chen, Ping-Che Yang, and Richard Tzong-Han Tsai. 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, 2018.

[26] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

[27] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[28] Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21273–21282, 2022.

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

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

[31] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[32] Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. Visual question generation as dual task of visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6116–6124, 2018.

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

[34] Xiao Lin and Devi Parikh. 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, 2016.

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

[36] Yang Liu, Guanbin Li, and Liang Lin. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

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

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

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

[40] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, July 2002. Association for Computational Linguistics.

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

[42] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark,

et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[44] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[45] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[46] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[47] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[48] Anna Rogers, Matt Gardner, and Isabelle Augenstein. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Computing Surveys (CSUR)*, 2022.

[49] Mourad Sarrouti, Asma Ben Abacha, and Dina Demner-Fushman. Visual question generation from radiology images. In *Proceedings of the First Workshop on Advances in Language and Vision Research*, pages 12–18, Online, July 2020. Association for Computational Linguistics.

[50] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July 2020. Association for Computational Linguistics.

[51] Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada. Customized image narrative generation via interactive visual question generation and answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8925–8933, 2018.

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

[53] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 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, 2022.

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

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

[57] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

[58] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

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

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

[61] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. Video as conditional graph hierarchy for multi-granular question answering. In

Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 2804–2812, 2022.

[62] Junbin Xiao, Pan Zhou, Angela Yao, Yicong Li, Richang Hong, Shuicheng Yan, and Tat-Seng Chua. Contrastive video question answering via video graph transformer. *arXiv preprint arXiv:2302.13668*, 2023.

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

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

[65] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 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, 2016.

[66] Min-Hsuan Yeh, Vincent Chen, Ting-Hao Huang, and Lun-Wei Ku. 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, December 2022. Association for Computational Linguistics.

[67] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019.

[68] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817, 2019.

[69] Zhenjie Zhao, Yufang Hou, Dakuo Wang, Mo Yu, Chengzhong Liu, and Xiaojuan Ma. Educational question generation of children storybooks via ques-

tion type distribution learning and event-centric summarization. *arXiv preprint arXiv:2203.14187*, 2022.

[70] Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: datasets, algorithms and challenges. *arXiv preprint arXiv:2203.01225*, 2022.

Appendix A

First appendix

A.1 Analyse Appendix

model	C G precision	C G recall	C G F1-score	T G precision	T G recall	T G F1-score
HGA[17]	**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 A.1: Baseline Model Evaluation Performance in Causal and Temporal Inference. C G represents the causal grounding metric. T G represents the Temporal causal 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 A.2: 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.

model	C G precision	C G recall	C G F1-score	T G precision	T G recall	T G F1-score
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 A.3: Evaluation performance of different sizes of T5 models with the two-stage tuning method in causal and temporal inference. C G represents the causal grounding metric. T G represents the Temporal causal 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**
Video 16to5 MLP	0.3028	0.3014	0.3021	0.3589	0.3316	0.3447

Table A.4: Vision Projection Matrix Evaluation Performance in Causal and Temporal Inference. C G represents the causal grounding metric. T G represents the Temporal causal grounding metric.

model	NN	WRB	VBD	VBZ	VB	JJ	VBG	WP	PRP
Video MLP	**4199**	**2692**	**1121**	1154	713	**504**	248	1038	220
Global Frame Comparison	4166	2571	981	**1489**	**776**	503	**345**	**1131**	**247**

Table A.5: Number of matching overlaps for various word types based on Spacy about the frame comparison 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
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 A.6: Global Frame Comparison Performance in Causal and Temporal Inference. C G represents the causal grounding metric. T G represents the Temporal causal grounding metric.

model	NN	WRB	VBD	VBZ	VB	JJ	VBG	WP	PRP
Global Frame Comparison	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

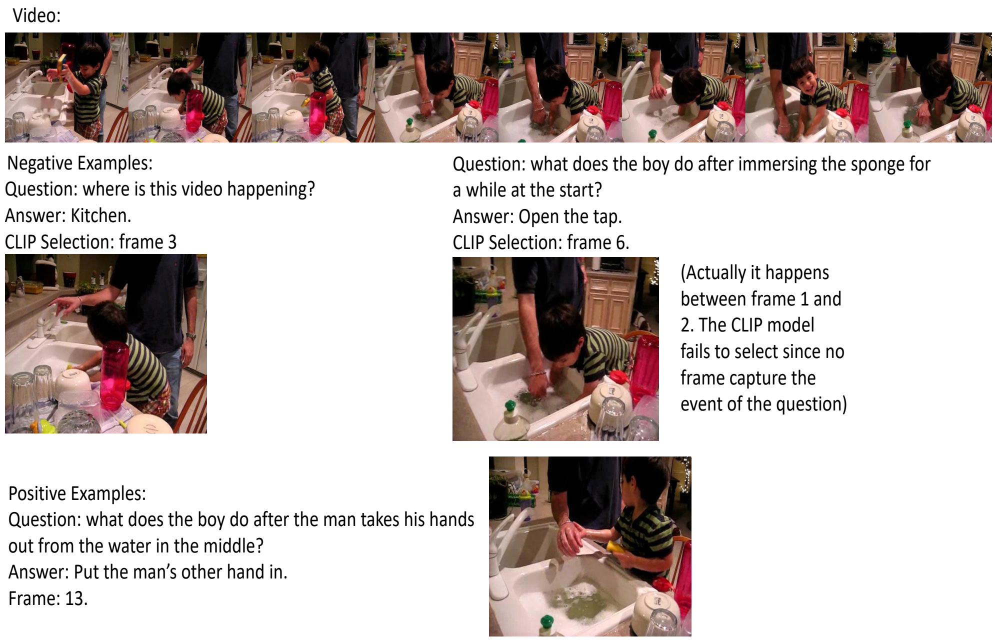
Table A.7: Number of matching overlap for various word types based on Spacy about the frame contrasting methods. NN means noun, singular or mass, WRB means wh-adverb, VBZ means verb, 3rd person singular present, VBD means verb, past tense, VB means verb, base form, JJ means adjective, VBG means verb, 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 A.8: 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	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 A.9: 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.


Video:

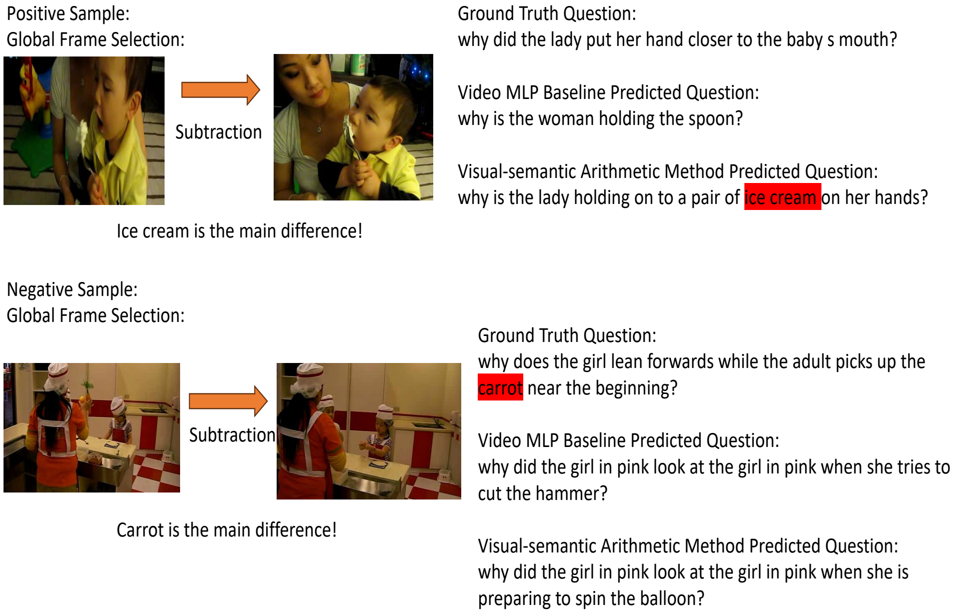
Negative Examples:
Question: where is this video happening?
Answer: Kitchen.
CLIP Selection: frame 3

Question: what does the boy do after immersing the sponge for a while at the start?
Answer: Open the tap.
CLIP Selection: frame 6.

(Actually it happens between frame 1 and 2. The CLIP model fails to select since no frame capture the event of the question)

Positive Examples:
Question: what does the boy do after the man takes his hands out from the water in the middle?
Answer: Put the man’s other hand in.
Frame: 13.

Figure A.1: CLIP Selection Performance



Positive Sample:
Global Frame Selection:

Ice cream is the main difference!

Ground Truth Question:
why did the lady put her hand closer to the baby s mouth?

Video MLP Baseline Predicted Question:
why is the woman holding the spoon?

Visual-semantic Arithmetic Method Predicted Question:
why is the lady holding on to a pair of ice cream on her hands?

Negative Sample:
Global Frame Selection:

Carrot is the main difference!

Ground Truth Question:
why does the girl lean forwards while the adult picks up the carrot near the beginning?

Video MLP Baseline Predicted Question:
why did the girl in pink look at the girl in pink when she tries to cut the hammer?

Visual-semantic Arithmetic Method Predicted Question:
why did the girl in pink look at the girl in pink when she is preparing to spin the balloon?

Figure A.2: The effectiveness of the Visual-semantic arithmetic method: check if the language model could recognize the difference between two frames.

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?

T5 small two-stage: (***56 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 on the sofa?",
“3”: "where is this video taken?",
“4”: "how did the girl hold the baby s hand?",
“5”: "why did the man in black hold the girl s hand?",
“6”: "what did the girl do after she sat down on the sofa?",
“7”: "what did the girl do after she sat down on the sofa?",
“8”: "why did the girl put her leg on the sofa after she sat down?",
“9”: "why did the man in black point to the girl s hand when she is sitting on the sofa?",
"10": "what did the man do after he sat down on the sofa?",
"11": "how did the man ensured he can see clearly?",
“12”: "why did the man laugh when the girl is talking?",
“13”: "why did the man in black hold the girl s hand?"

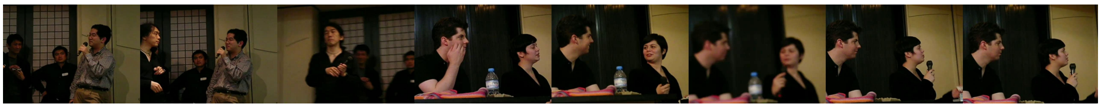
T5 Large two-stage: (***58 matching overlap***)
“1”: "how did the man keep his hair out of his face?",
“2”: "what did the girl do after the man touched her face?",
“3”: "where is this video taken?",
“4”: "how did the man hold the girl s hand?",
“5”: "why did the man in black touch the girl s face?",
“6”: "what did the man do after the girl sat down?",
“7”: "what did the girl do after the man touched her face?",
“8”: "why did the man in black touch the girl s leg after the girl sat down?",
“9”: "why did the man in green point to the girl s seat at the end of the video?",
"10": "what did the man do after the girl sat down?",
"11": "how did the man ensure he can see clearly?",
“12”: "why did the man laugh when the girl is talking?",
“13”: "why did the man in black move his head backwards in the middle of the video?"

T5 small one-stage: (***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?"

T5 Large one-stage: (***44 matching overlap***)
“1”: "how did the girl kept her hair out of her face?",
“2”: "what did the girl do after she touched the man s hair?",
“3”: "where is this video taken?",
“4”: "how did the girl in white dress touched the cake at the start of the video?“,
“5”: "why did the lady in white move the girl s hands?",
“6”: "what does the girl do after standing for a while at the end?",
“7”: "what does the girl do after looking at her right in the middle?",
“8”: "why does the girl in pink stop her spinning after a while?",
“9”: "why did the lady in white point at the table at the end of the video?",
"10": "what does the boy do after looking at the lady for a while at the end?",
"11": "how did the lady see her surrounding clearly?",
“12”: "why did the man in black smiled after the girl lied on the table?",
“13”: "why does the man in black move backwards in the middle of the video?"

Figure A.3: language Size Performance. Yellow scopes represent matching overlap with ground truth questions. Red words represent repetitive questions.

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 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?"

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?"

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 A.4: 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.